IEEE Access
The journal for rapid open access publishing

IEEE Xplore Digital Library
http://ieeexplore.ieee.org/Xplore

Using Sequential Decision Making to Improve Lung Cancer Screening Performance

PANAYIOTIS PETOUSIS[1], AUDREY WINTER[2], WILLIAM SPEIER[2], DENISE R. ABERLE[1,2], WILLIAM HSU[1,2], ALEX A.T. BUI.[1,2],

***Abstract*—Globally, lung cancer is responsible for nearly one in five cancer deaths. The National Lung Screening Trial (NLST) demonstrated the efficacy of low-dose computed tomography (LDCT) to identify early-stage disease, setting the basis for widespread implementation of lung cancer screening programs. However, the specificity of LDCT lung cancer screening is suboptimal, with a significant false positive rate. Representing this imaging-based screening process as a sequential decision making problem, we combined multiple machine learning-based methods to learn a partially-observable Markov decision process that simultaneously optimizes lung cancer detection while enhancing test specificity. Using NLST data, we trained a dynamic Bayesian network as an observational model and used inverse reinforcement learning to discover a rewards function based on experts' decisions. Our resultant predictive model decreased the false positive rate while maintaining a high true positive rate at a level comparable to human experts. Our model also detected a number of lung cancers earlier.**

***Index Terms*—Early disease prediction, Dynamic Bayesian networks, Lung cancer screening, Partially observable Markov decision processes, QMDP algorithm**

I. Introduction

Lung cancer is the leading cause of cancer-related mortality, estimated to be responsible for 2.1 million deaths worldwide in 2018. Although the five-year survival rate for this disease improves when discovered in its nascent stages [1], only 15% of all lung cancers are detected early as symptoms often do not appear until the disease has advanced to a late or terminal stage. The findings of the National Lung Screening Trial (NLST) and more recent NELSON study [2], [3] support the implementation of lung cancer screening programs to identify individuals at high-risk for developing this disease, using low-dose computed tomography (LDCT) imaging to maximize initial detection [2], [4]. Lung cancer screening guidelines consider the high-risk population as a whole, balancing the benefit of longitudinal observation of pulmonary nodules against pragmatic issues including test sensitivity and specificity. Of current concern is the disproportionately high false positive (FP) rate for LDCT screening: in the NLST, the overall positive screen rate with LDCT was 24%, yet the positive predictive value of a positive screen was less than 4% [5]. Of the total number of lung nodules diagnosed in the NLST, only 3-6% were found to be malignant, depending on nodule size. The negative consequences of overdetection are significant, increasing the use of unnecessary diagnostic procedures results in complications and patient duress [6]. As such, in this work we present a novel approach that reduces the FP rate associated with LDCT lung cancer screening while maintaining a high true lung cancer detection rate.

Applying computational methods on the growing amount of data available from electronic health records (EHRs) and imaging in this domain, we can address such issues and begin to ask the more precise question of how to optimize lung cancer screening for each person, discovering better ways to risk-stratify as an individual's disease trajectory unfolds. But individually-tailoring this process is not straightforward: the obvious next "best" action may not be ideal in the long run given a patient's evolving risk factors, potential future observations, and changing benefit of decisions as time progresses. Sequential decision making, a class of algorithms used in artificial intelligence (AI) for selecting the series of actions optimizing the likelihood of achieving a goal in dynamic environments, provides one way to overcome this difficulty. Known as the temporal credit assignment problem [7], we pose lung cancer screening in terms of finding the series of actions, given various observations over time, which maximizes early disease detection while minimizing false positives.

We developed a predictive model informing personalized lung cancer screening policies using machine learning and sequential decision making methods. Specifically, we established a framework for learning a partially-observable Markov decision process (POMDP), progressively optimizing the choice of screening actions given prior observations. POMDPs, a generalization of Markov decision processes, are useful when serial observations of a disease are indirect and/or subject to interpretation. Demonstrated in other settings [8]–[11], the implementation of POMDPs to guide clinical decision making is challenging given the need to derive required probability distributions and reward functions. We leveraged different techniques to learn a POMDP from NLST data: we integrated a dynamic Bayesian network (DBN) into the POMDP to predict the chance of developing lung cancer and to determine the POMDP's observation and transition probabilities, and we applied inverse reinforcement learning (IRL) to formulate a rewards model [12], mimicking experts' decisions.

We trained and tested our POMDP using a dataset of 5,402 single nodule unique trajectories of lung cancer screening patients from the NLST LDCT trial arm. We compared our model's decisions with experts' decisions over time, and found that: 1) our POMDP lowered the false positive rate for most screenings in the NLST, while maintaining true positive detection rates; and 2) our POMDP improves early prediction of cancer cases with indeterminate pulmonary nodules (IPNs, nodules having some risk of developing into cancer [13]) as

Digital Object Identifier: 10.1109/ACCESS.2019.2935763

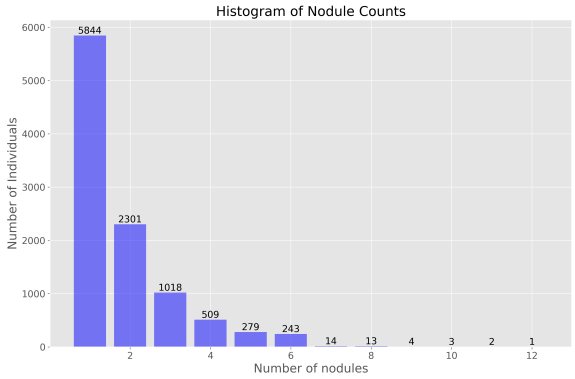


Fig. 1: Histogram of nodule counts per NLST subject.

compared to radiologists' interpretation.

II. Methods

A. NLST dataset

The NLST was a multi-site randomized controlled trial examining the impact of two imaging modalities, LDCT and chest x-ray, for early detection of lung cancer in asymptomatic, high-risk individuals. Over 53,000 participants underwent three annual screenings with follow-up to assess cancer outcomes. Participants suspected of cancer were referred for diagnostic procedures (e.g., biopsy) and removed from the study for treatment if lung cancer was confirmed. In this work, we used data gathered from NLST's LDCT arm. The dataset comprises over 25,000 participants with information on demographic, clinical, and imaging data. Of this population, only 10,231 cases had one or more solitary IPNs over the study period. Figure 1 shows the number of patients and total number of nodules reported. Unfortunately, NLST annotation data did not uniquely identify individual nodules in participants with multiple nodules, making linking observations over serial scans difficult. As such, we further constrained our data to individuals with only one IPN reported in the same anatomical lung lobe during the study, assuming that the same nodule was observed over time. This selection criteria and preprocessing to remove inconsistent cases (see below) resulted in a total of 5,402 cases, which we used to train and test our POMDP model.

To perform a five-fold stratified cross validation ($80:20\%$ training:test ratio) with this data, we randomly generated each fold while maintaining the relative proportion of cancer to non-cancer cases seen at each screening time point of the NLST study.

B. Data preprocessing

Table I summarizes the NLST variables used in our analysis. We considered the same demographic and clinical variables selected in the Tammemägi model [14] and replicated its preprocessing steps. We converted two variables into binary representations: family history of lung cancer (if any first degree relative had a history of lung cancer) and personal lung cancer history (if the individual had any prior history of lung cancer). Missing values for the variables used with the Tammemägi model were imputed using a variation of a multiple clustering imputation approach [15]. In addition to the radiologist's overall interpretation of the LDCT scan, we employed several imaging features describing the nodule, discretizing continuous variables: location, nodule size (Bin 1: ≥ 0 mm and $\leqslant 3$ mm; Bins 2-9: 1 mm bins from $3-11$ mm; Bin 10: >11 mm and $\leqslant 27$ mm; and Bin 11: >27 mm), predominant attenuation, and margins. Given the sparsity of cases with nodules of size >11 mm, we created larger bins by identifying discretizations maximizing POMDP performance using the training data. We removed inconsistent cases with a perpendicular measurement greater than the reported longest nodule diameter and any cases with missing measurements. Cases without screening abnormalities at an annual screening for reasons other than death, cancer, or missed screening, a nodule size between 0-3 mm was assumed. Cases without nodule size abnormalities across the three annual screenings were excluded from the analysis. Nodule size was then interpolated between annual screenings using the average value between time points, with nodule consistency, margins, and follow-up decisions unchanged relative to the earlier annual observation. This interpolation, used only when training (learning the model's parameters), augmented temporal data points every six month intervals for the training data and improved the overall performance of our POMDP model when testing. Other variables used in the model include the total number of screening days, occurrence of diagnostic procedures (biopsy, thoracotomy, diagnostic CT exam), and confirmed diagnoses of lung cancer.

C. Defining and learning the POMDP components

1) States (s) and actions (a): Figure 2(a) illustrates the lung cancer screening POMDP state space, observations, and potential state transitions. We adopted a state space used in our earlier work [16]. This state space consists of three states defined around the true cancer state of each subject after each screening. No-cancer (NC) is the state in which the individual has no remarkable findings for lung cancer (e.g., nodules <4 mm). The Uncertain (U) state is an intermediate state in which an individual exhibits suspicious abnormalities (e.g., lung nodules ≥ 4 mm) but no confirmed diagnosis of lung cancer. The Lung Cancer (LC) state represents any case with a confirmed lung cancer diagnosis through the use of additional procedures. LC is a terminal state in which an individual enters and simultaneously leaves the screening process (as NLST participants diagnosed with lung cancer were removed from the clinical trial for treatment). We simplified the set of possible actions into two types, embodying the core decisions made by experts: to continue screening with a follow-up LDCT or to recommend an intervention (i.e., any procedure performed in relation to diagnosing lung cancer).

2) Observations (z): Following from the NLST's screening paradigm, two types of observations are possible: those coming from annual screens (LDCT findings) and interpretation and those arising from a diagnostic intervention. To capture

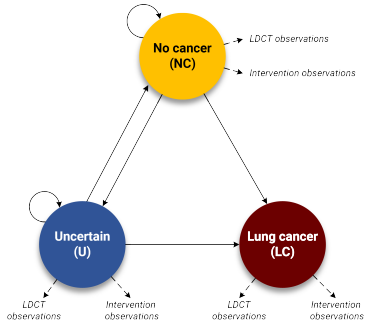
Demographic, clinical, outcome variables (% missing)	Variable type	Value	Mean (SD)/Category proportions (%)
Age (0%)	Continuous		61.64 (5.05) years
Education (0.31%)	Categorical	8th grade or less	1.34%
		$9-11^{th}$ grade	4.66%
		High school graduate/GED	24.29%
		Post-high school training, excluding college	13.72%
		Associate degree/some college	23.53%
		Bachelors degree	16.47%
		Graduate/professional school	14.09%
		Other	1.89%
Race (1.48%)	Categorical	White	93%
		Black	4.28%
		Asian	2.05%
		American Indian or Alaskan Native	0.23%
		Native Hawaiian or Other Pacific Islander	0.26%
Body mass index (0.03%)	Continuous		27.61 (4.92)
Chronic obstructive pulmonary disease (COPD) (0.26%)	Binary	No	94.38%
		Yes	5.62%
Family history of lung cancer (0%)	Binary	No	77.32%
		Yes	22.68%
Personal history of lung cancer (0%)	Binary	No	95.59%
		Yes	4.41%
Smoking status (0%)	Binary	No	50.19%
		Yes	49.81%
Smoking intensity (0%)	Continuous		28.71 (11.43)
Duration of smoking (0%)	Continuous		40.2 (7.27) years
Smoking quit time (0.48%)	Continuous		3.67 (4.95) years
Confirmed lung cancer diagnosis (0%)	Binary	No	91.65%
		Yes	8.35%

Study variable	Variable type	Value	t_0	t_1	t_2
Screening outcome (radiologist interpretation)	Categorical	Negative screen, no significant abnormalities	8.83%	4.59%	4.15%
		Negative screen, minor abnormalities not suspicious for lung cancer	24.07%	25.84%	49.09%
		Negative screen, significant abnormalities not suspicious for lung cancer	6.42%	3.74%	4.41%
		Positive, change unspecified, nodule(s) $\geq$ 4 mm or enlarging nodule(s)	60.16%	7.09%	0%
		Positive, no significant change, stable abnormalities	0%	36.82%	17.66%
		Positive, other	0%	13.44%	13.35%
		Not compliant, left study	0%	0.39%	0.68%
		Not compliant, refused a screen	0.35%	4.46%	4.92%
		Not compliant, wrong screen	0.15%	0%	0.02%
		Not compliant, erroneous report of lung cancer before screen (LSS only)	0%	0%	0.04%
		Not compliant, form not submitted, window closed	0%	0.07%	0.11%
		Not expected, cancer before screening window	0%	2.78%	4.09%
		Not expected, cancer in screening window	0%	0.04%	0.15%
		Not expected, death before screening window	0%	0.46%	1.07%
		Not expected, death in screening window	0%	0.28%	0.26%

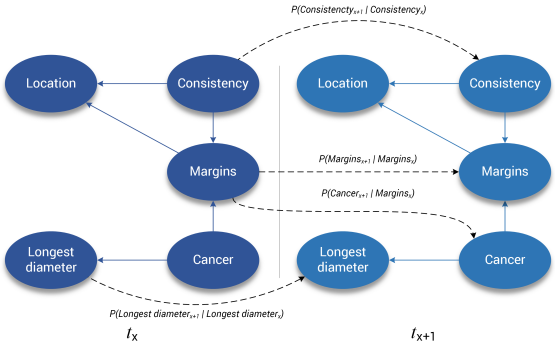
Nodule variables (% missing)	Variable type	Value	t_0	t_1	t_2
Location (t_0 =40.00%,t_1 =41.58%,t_2 =40.93%)	Categorical	Right upper lobe	24.44%	23.92%	22.78%
		Right middle lobe	13.21%	13.97%	13.22%
		Right lower lobe	23.60%	23.89%	24.16%
		Left upper lobe	13.79%	12.77%	13.51%
		Lingula	4.01%	3.64%	3.79%
		Left lower lobe	20.58%	21.1%	21.53%
		Other	0.37%	0.70%	1.00%
Margins (t_0 =40.00%,t_1 =40.30%,t_2 =40.93%)	Categorical	Spiculated (stellate)	12.5%	9.05%	7.62%
		Smooth	62.97%	67.97%	70.79%
		Poorly defined	18.54%	19.01%	18.8%
		Unable to determine	5.99%	3.97%	2.79%
Longest diameter (t_0 =40.00%,t_1 =41.58%,t_2 =40.93%)	Continuous		7.97 (7.00) mm	7.23 (5.19) mm	7.07 (5.50) mm

Diagnostic intervention variables	Variable type	Value	t_0	t_1	t_2
Biopsy	Binary	No	95.17%	97.6%	97.28%
		Yes	4.83%	2.37%	2.72%
Invasive procedure	Binary	No	94.95%	97.57%	97.17%
		Yes	5.05%	2.43%	2.83%
Non-invasive procedure	Binary	No	46.91%	68.72%	80.91%
		Yes	53.09%	31.28%	19.09%

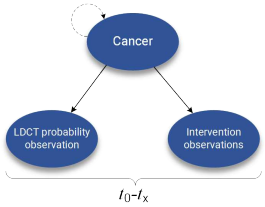
TABLE I: Variables used for the development and evaluation of the lung cancer screening POMDP model. After applying selection criteria and removing subjects with inconsistent values, a total of 5,402 LDCT screening cases were used from the NLST. Percentages of missing data are provided, alongside categorical breakdowns and mean values.



(a) Lung cancer screening POMDP.



(b) DBN-based observation model.



(c) Transition model.

Fig. 2: (a) NC represents a non-cancer state, U is an intermediate uncertain cancer state, and LC is the lung cancer state. Arrows indicate allowed transitions between states. LDCT and intervention observations represent the possible observations of the model in each state. **(b)** The nodule size node represents the possible categories of nodule size. The consistency node represents the categories of nodule consistency and the margin node the categories of nodule margins. The Cancer node represents the categories of cancer or no cancer. t_0 represents the intra-slice structure of the model. Solid line arrows represent the intra-slice interactions between nodes. The inter-slice structure is depicted between the t_0 and t_1 time slices. Dashed arrows represent inter-slice interactions between variables over time. This DBN is recurring for 5-time steps ($x = 4$). **(c)** The LDCT probability observations represent the 100 bins of probabilities as categories. The Intervention observations node consists of two categories the observation of cancer or not, from diagnostic procedures. The Cancer node consists of three states the NC, U and LC cancer states.

the interactions between the nodule size, consistency, and margins we used a model to combine the observations into a single representation as a probability. Specifically, we used a DBN to infer the probability of cancer over time from these observations. Alternative models were considered, including logistic regression, and an exhaustive search of all combinations of observations, with the DBN and the exhaustive search demonstrating the best performance in conjunction with the POMDP (see Appendix Table XVII). The DBN topology was learned from the data: we learned the intra-slice structure of the DBN (i.e., conditional dependencies between variables in the same time step) using t_0 observations from the K2 algorithm in the Bayes Net Toolbox (BNT) [17]; and inter-slice structure (i.e., dependencies over time) was learned using cases that had a complete trajectory of screening over the NLST screening period (i.e., no missing observations) using the batch Expectation-Maximization (EM) algorithm also in BNT. Figure 2(b) shows the intra- and inter-slice structure of the learned model, which we then parameterized using training data. In the POMDP, we then used this DBN with observations of a given patient to infer a probability of cancer over time as our new observation. These probabilistic observations were discretized in 100 equal sized bins, from 0-1. For intervention observations, we determined if an individual undergo an intervention and was diagnosed with cancer or did not undergo an intervention.

3) Transition and observation probabilities: Transition and observation probabilities were computed using a dynamic Bayesian network, per Figure 2(c): the LDCT node represents a conditional probability table (CPT) of 100 categories corresponding to each discretized probability; the Intervention node represents a CPT table of two observations, cancer after an intervention or no cancer with or without an intervention; and the Cancer node represents a CPT table of three categories per our state model. Usually, the transition probabilities of a POMDP are different based on the choice of action in a given state ($T(s_j, s_i, a)$). The transition matrix used for the lung cancer POMDP model is assumed to be invariant of action. But the observation matrix ($O(z, s, a)$) is state and action dependent. We modeled the observations of Intervention as being impossible (i.e., probability of zero) when the action of LDCT is performed and the observation of an LDCT as impossible when the action of Intervention is performed. An important implementation note is in regards to sparsity, as some LDCT probabilities will be calculated as zero given no instances in the dataset (although they are feasible in real-world settings). Thus, to deal with sparsity we replaced all zero probabilities with a very small probability (0.0001) and normalized over the matrix to improve overall inference [18].

4) Rewards: A POMDP's reward function defines the behavior of the agent as it aims to optimize based on returned values. In our POMDP, we define rewards in terms of a state-action pair (R(s,a)). We learned a reward function using the recommendations of experts from the NLST dataset. Using inverse reinforcement learning (IRL), we learned state and action rewards via an adaptive maximum entropy IRL algorithm [12]. A multiplicative model was then employed to learn each

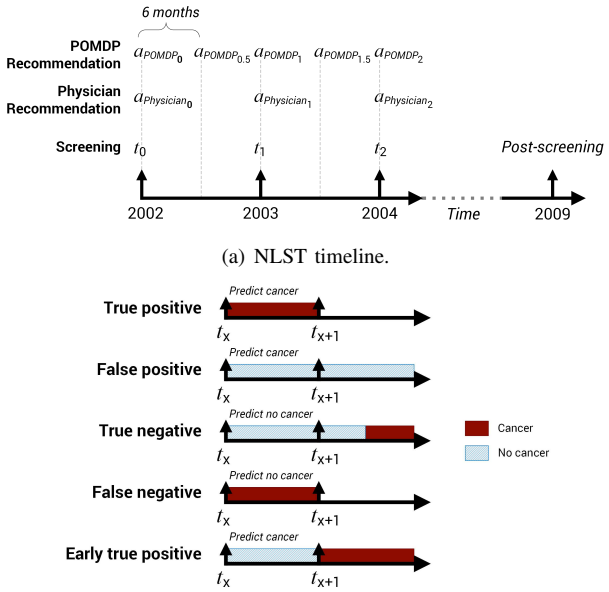
combination of state-action pair rewards.

5) Initial beliefs: In a POMDP, the belief state is a probability distribution over the states of the process. The initial belief is the initial probability distribution over the states at time t_0. To generate initial beliefs for each individual we used the $\text{PLCO}_{\text{M2012}}$ model [14] with demographic and clinical features at baseline to predict the risk of cancer. Tammemägi et al. [19] used the Prostate, Lung, Colorectal and Ovarian Cancer (PLCO) Screening Trial to develop 6-year lung cancer risk models. The models achieved high discrimination and calibration performance. The $\text{PLCO}_{\text{M2012}}$ is an updated version of the original model trained and validated on the PLCO dataset and externally validated on the NLST cohort. The variables and weights of the logistic regression model used are the same as reported in the $\text{PLCO}_{\text{M2012}}$ model [14]. Demographic features include age, education, race, and body mass index (BMI). Clinical features encompassed the presence of chronic obstructive pulmonary disease (COPD), family history of lung cancer, personal history of cancer, smoking status, smoking intensity, and duration of smoking. To generate an initial belief of cancer over the three states of our state space, we used the following rule: the probability of the LC state is the risk of cancer times two computed by the $\text{PLCO}_{\text{M2012}}$ model; the probability of the U state is assumed to be zero and the probability of the NC state is the complement of LC. To update beliefs we follow the basic recursive filtering rule [20], given by Equation 1 where α is a normalization constant such that $\alpha = \frac{1}{\sum_{s_j} P(o|s_j) \sum_{s_i} P(s_j|s_i,a)b(s_i)}$.

$$b'(s_j) = \alpha \cdot P(o|s_j) \sum_{s_j} P(s_j|s_i, a)b(s_i) \tag{1}$$

D. Solving the POMDP model

POMDP models can be solved through the value iteration (VI) algorithm. However, the number of possibilities to be considered is exponential in terms of the number of states, actions, and observations modeled. At each time step, the VI algorithm enumerates k^{Ω} new policies trees, where k is the previous time step number of policy trees and Ω is the number of observations. Each policy tree represents a linear function. For an infinite horizon process the value function will have infinite linear functions, a key reason why POMDPs are often considered impractical. To solve infinite horizon problems, we can use approximation algorithms [20]–[23], providing significant speed-up. Markedly, our proposed POMDP model has three states, two actions and 102 observations. We opted to use the QMDP approximation algorithm, shown in Algorithm 1 in the Appendix. QMDP solves the POMDP problem as an MDP and then generalizes the value function into a POMDP. More specifically, the QMDP algorithm estimates the value function for the equivalent MDP while ignoring the observation model. The MDP value function is used to define a linear function (i.e., a policy tree) for each action. The main disadvantage of the QMDP algorithm is that it dismisses the state uncertainty described in POMDPs but solves the POMDP with MDP computational time complexity. To select optimal actions that maximize expected utility we use Algorithm 2 in the Appendix, which given a belief and the Q matrix computes their dot product to compute the utility of each action when being in a belief (b).



(a) NLST timeline.

(b) Definition of evaluation metrics.

Fig. 3: (a) Screenings represent annual LDCT imaging observations with information about the subject's cancer status. In contrast, our POMDP model suggests screening recommendations every six months. **(b)** Illustration depicting true positive/negative and false positive/negative cases for the POMDP's performance over time. The colored bars indicate truth based on the NLST observations and subjects' known outcomes. We also demonstrate how early true positives are defined in this study.

III. Results

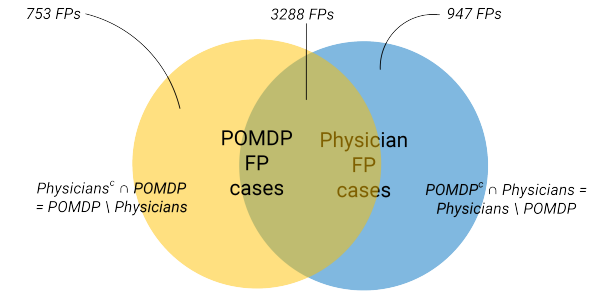
NLST participants underwent three annual screenings with follow-up over six years to identify subsequent lung cancers. At each screening time point (t_0, t_1, t_2), a radiologist read the imaging study and made a decision to refer patients for a diagnostic procedure (e.g., early repeat LDCT, diagnostic CT, PET-CT, or biopsy/tissue sampling) or to continue annual LDCT screening. Our POMDP suggests actions at these three screening time points as well as between the screenings using imputation, resulting in five recommendations in 6-month intervals (Figure 3(a)). Observations used by the POMDP include imaging features about nodule size, margins, location, and consistency. Our evaluation examines the POMDP's recommended actions over all five points $(a_{POMDP_0}, a_{POMDP_{0.5}}, a_{POMDP_1}, a_{POMDP_{1.5}}, a_{POMDP_2})$ and directly compares against the physicians' performance at the annual screenings.

A. POMDP versus physician performance

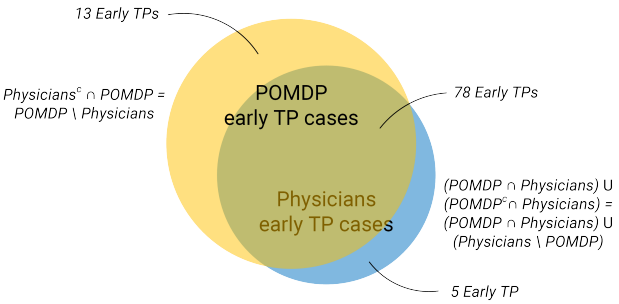
To compare the performance of the POMDP model against physicians we calculated the precision (positive predictive value, PPV), recall/true positive (TP) rate (sensitivity), and true negative (TN) rate (specificity) for recommended actions at each screening point. We used the following criteria to assess

our model: if the POMDP suggests a diagnostic intervention and the individual is subsequently diagnosed with cancer in the following time period, it is counted as a true positive, otherwise it is considered a false positive; if the POMDP suggests no diagnostic intervention, but an annual LDCT screen, and the individual is diagnosed with cancer in the following screen, it is a false negative (FN), otherwise it is a true negative (Figure 3(b)).

We assessed our POMDP's performance based on a five-fold cross validation design. To match physicians' TP rates (who had a lower threshold for positive screens) and obtain comparable results, we adjusted the POMDP rewards function (using the training data) to be more conservative. We then evaluated this updated POMDP on our testing data. Table II shows the results of the POMDP model with tuned rewards against physicians' performance. Our model reduces the FP rate in most screenings (t_1, t_2, and post-screening) compared to the experts while maintaining a high TP rate for screening: at t_0, TN and TP rates are 2% lower and 3% lower than the physicians'; at t_1, TN and TP rate are 1% higher and 3% higher; at t_2, TN and TP rate are 4% higher and 4% lower; and in the post-screening period the POMDP's TN and TP rate are 3% higher and 8% higher than the experts', respectively. We also analyzed the performance of the POMDP model for earlier cancer detection (i.e., detection of a t_2 cancer at t_1). The detection of early TPs is also improved with earlier diagnostic recommendations (e.g., the TP rate for action $a_{POMDP_0}, a_{POMDP_{0.5}}, a_{POMDP_1}, a_{POMDP_{1.5}}$ for t_2 and post-screening) compared to physicians' recommendations. The POMDP TP rate is higher than the physician's over time for post-screening, as depicted in Figure 5 and discussed in the following section.



(a) False positive cases between POMDP and physicians.



(b) Early predicted positive cases between POMDP and physicians.

Fig. 4: Comparison of case agreement between the POMDP and experts. The numbers in each subset represent the total number of FPs or early TPs grouped from every testing set for each fold of the five folds. **(a)** Yellow: Cases predicted as false positives by the POMDP model. Blue: Cases predicted as false positives by the physicians. The union of these groups are all cases predicted by the POMDP or physicians as false positives. POMDP^c represents the complement of the POMDP set. **(b)** Yellow: Cases predicted as early true positives by the POMDP model. Blue: Cases predicted as early true positives by the physicians.

B. Understanding POMDP and physician differences

We calculated a kappa score to test the level of agreement between physicians and the POMDP. Notably, kappa values trended lower, implying that the POMDP and experts classify different cases positively over time, which influences the FP rate. To elucidate this difference, we grouped subjects predicted to have lung cancer by the POMDP vs. physicians, analyzing cases where they had different predictions. The preponderance of subjects different between the groups were individuals classified as FPs or early TP cases (i.e., cases predicted as positives earlier by the POMDP relative to their cancer diagnosis in the NLST trial). Figure 4 depicts these two cohorts. We explored the feature distributions of each group to assess similarity. We used chi-squared or Fisher's tests for categorical variables and the Student or Wilcoxon-Mann-Whitney tests for continuous variables. Additionally, to assess the effect size of the computed p-value we used the Cramer's V and the r^2 or Cohen's r^2 effect size, correspondingly, for each test [24]–[26]. Tests with p-values < 0.05 were considered significant. The false positive analysis showed that smoking years, age, largest nodule size at t_0, and smoking quit time had significantly different distributions and the largest effect size between the groups of post-screening cases (see Table III). The additional early prediction TP cases predicted by the POMDP model in comparison with the physicians showed that nodule size at t_0 (largest nodule diameter) and smoking years were significantly different between the groups. The nodule size at t_0 was shorter and years of smoking less than the early TPs predicted by both the physicians and POMDPs (see Table III and Figure 4). A full analysis comparing these groups is presented in the Appendix.

C. POMDP stability

In the NLST, a minimum threshold of 4 mm was used to classify findings as nodules. A later analysis [27], [28] showed that changing this threshold to 6 mm significantly reduced the FP rate while maintaining the same TP rate [29], [30]. As such, we stratified our cases into nodules < 6 and ≥ 6 mm at baseline and tested the POMDP. To assess the robustness and performance distribution of the POMDP model we performed a bootstrap evaluation, randomly sampling from our NLST dataset 240 times to define our training and testing sets. Subsequently, all performance measures for each seed were used to calculate the median, the interquartile range (IQR), and the range for each metric. This analysis is summarized in Figure 5, where the box plots depict the median and IQR of each action. Significance tests were performed using the Wilcoxon signed rank test or paired t-test as appropriate to assess if the performance distribution of the POMDP is

	Cancers	Non-cancers		POMDP			Physicians			Kappa
				TN rate	Recall	Precision	TN	Recall	Precision	
Screening t_0	32	1,047	a_0	0.46	0.97	0.05	0.48	1.00	0.06	0.41
Screening t_1	17	1,030	a_0	0.47	0.67	0.02	0.48	0.39	0.01	0.40
			$a_{0.5}$	0.46	0.67	0.02				
			a_1	0.34	0.98	0.02	0.33	0.95	0.02	0.28
Screening t_2	21	1,009	a_0	0.47	0.56	0.02	0.48	0.28	0.01	0.40
			$a_{0.5}$	0.47	0.56	0.02				
			a_1	0.35	0.70	0.02	0.32	0.46	0.01	0.27
			$a_{1.5}$	0.34	0.72	0.02				
			a_2	0.25	0.96	0.03	0.21	1.00	0.03	0.06
Post-screening	19	900	a_0	0.47	0.71	0.03	0.48	0.46	0.02	0.40
			$a_{0.5}$	0.47	0.71	0.03				
			a_1	0.35	0.82	0.02	0.32	0.71	0.02	0.25
			$a_{1.5}$	0.34	0.82	0.02				
			a_2	0.25	0.94	0.02	0.22	0.86	0.02	0.05

TABLE II: POMDP vs. physician performance, 5-fold cross validation using test data partition (average across runs presented). A kappa score was also calculated to compare the level of agreement between the model and experts.

Variables	**False positives analysis**		**Early true positives analysis**	
	POMDP	Physicians	POMDP	Physicians
Nodule size t_2 (mm)	**4.73**$^{+}$	**3.35**$^{+}$	5.17	5.24
Nodule size t_1 (mm)	3.8	3.61	3.17	5.79
Nodule size t_0 (mm)	**2.47**$^{++}$	**3.86**$^{++}$	**1.54**$^{+}$	**4.82**$^{+}$
Years of smoking	**41.64**$^{+++}$	**35.52**$^{+++}$	**40.31**$^{+}$	**45.33**$^{+}$
Years since quitting smoking	**2.79**$^{++}$	**5.21**$^{++}$	4.77	2.29
Age at baseline	**61.8**$^{++}$	**58.68**$^{++}$	62.92	64.89
Smoking status at baseline (% smokers)	**60.29**$^{+}$	**32.52**$^{+}$	46.15	65.38

TABLE III: Feature analysis of cases different between the POMDP and physicians, comparing false positives and early true positives. Reported values represent the post screening average values per variable. Bold values represent features with statistically significantly different distributions. The magnitude of the effect size of the p-value computed using the Cramer's r^2, the r^2, and the Cohen's r^2 are color-coded as: orange, small effect size ($^{+}$); blue, medium effect size ($^{++}$); and black, large effect size ($^{+++}$). The Cramer's r^2, the r^2, and the Cohen's r^2 ranges for small, medium, and large are given in the Appendix.

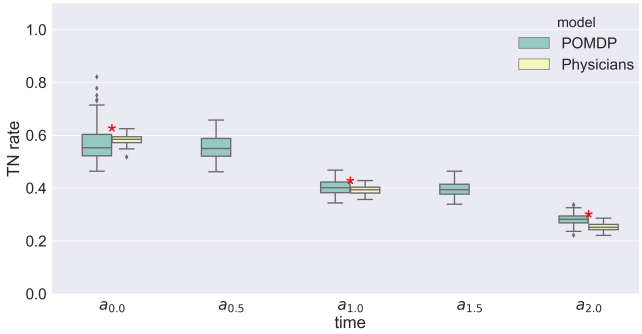
significantly different from that of physicians. Tests with p-values < 0.05 were deemed significant.

Interestingly, the POMDP model's CIs become narrower over time, suggesting that it stabilizes with longer trajectories of observations. When only testing the POMDP model on a cohort of cases with nodules larger than 6 mm at baseline, the POMDP model improves the true negative rate (i.e., reduces the FP rate) while maintaining a TP rate comparable to the physicians. Markedly, precision is significantly improved using the POMDP model in this scenario. When testing on the cases with nodules smaller than 6 mm at baseline, initially the POMDP TN rate is lower than that of physicians' but improves over time. The TP rate and early prediction of cancer is significantly improved compared with physicians in post-screening. Precision is also significantly improved for all screenings. This comparison of cases that are typically easier to classify as cancerous due to a lager nodule size (i.e., $\geq$ 6 mm) demonstrates how our approach reduces FPs associated with lung cancer screening. Additionally, in the situation where IPNs are smaller ($<$ 6 mm), our model still improves early prediction and overall precision. Box plots with the smaller than 6 mm and larger than 6 mm cohorts combined is presented in Figure 6 in the Appendix.

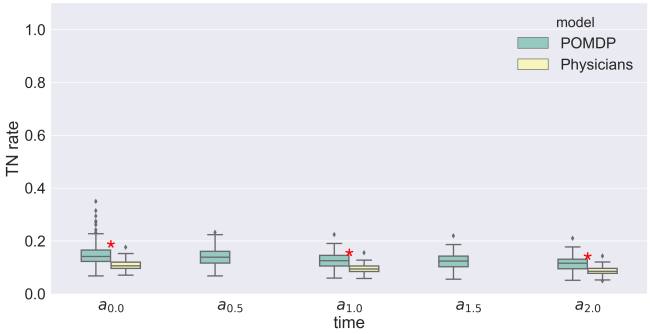
IV. Discussion

The majority of individuals diagnosed with lung cancer have a low 5-year survival rate of 18% [31]. In sharp contrast, earlier detection of this cancer improves this statistic threefold to 56% [31]. While LDCT lung cancer screening aims to reduce mortality through earlier detection, the FP rate associated with IPNs remains high, with concomitant concerns of increased healthcare costs and unnecessary psychological burden for patients. To address this concern, we developed a POMDP for lung cancer screening, demonstrating simultaneous reduction in FPs and earlier cancer detection when compared to experts' performance. Maintaining a high TP rate while minimizing the FP rate is challenging given the correlation of nodule malignancy and size: larger nodules tend to be malignant; and conversely, nodules smaller than 6 mm are less likely to be cancerous. We improved the TN rate for nodules larger than 6 mm at baseline while maintaining a true positive rate on par with experts. When comparing our POMDP against physicians' predictions for cases with nodules smaller than 6 mm, improved true positive rate and precision overall were seen, while progressively increasing the TN rate (see Figure 5).

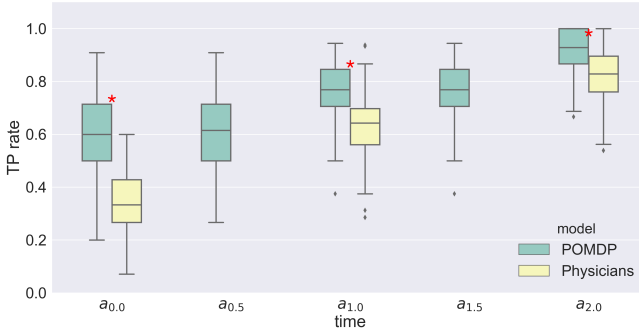
Our POMDP uses a DBN to generate observations for a patient over time that are used to update its belief about lung



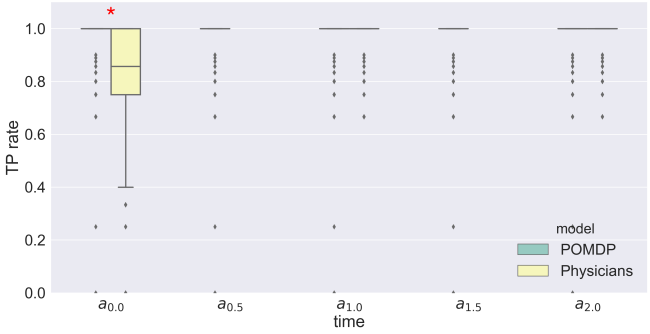
(a) TN rate on individuals with nodules smaller than 6 mm at baseline.



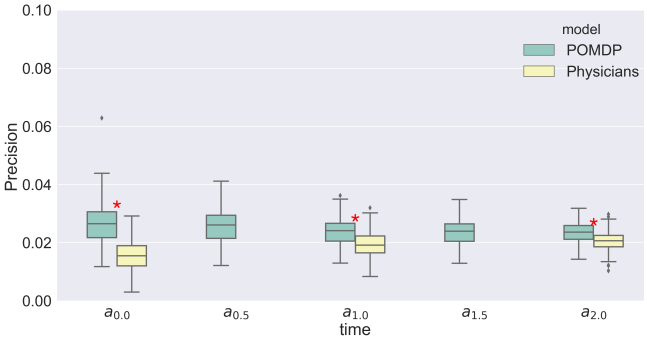
(b) TN rate on individuals with nodules larger than 6 mm at baseline.



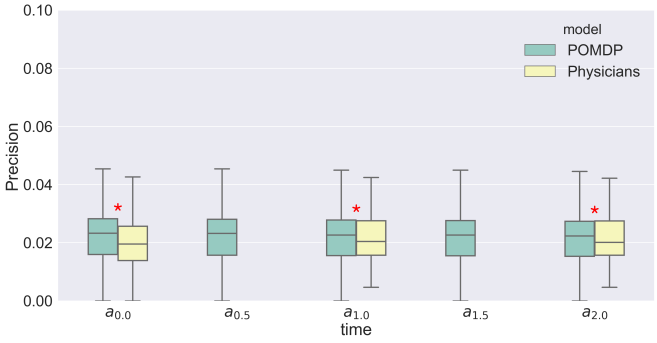
(c) Recall/TP rate on individuals with nodules smaller than 6 mm at baseline.



(d) Recall/TP rate on individuals with nodules larger than 6 mm at baseline.



(e) Precision on individuals with nodules smaller than 6 mm at baseline.



(f) Precision on individuals with nodules larger than 6 mm at baseline.

Fig. 5: Box plots of the performance (TN, TP, precision) of the POMDP and physicians on cases from the NLST testing set, from the start of the trial through to last screening. **Left column**: Cases with nodules smaller than $6mm$ at baseline. **Right column**: Cases with nodules larger than $6mm$ at baseline. Blue and yellow represent the POMDP and experts, respectively. Red stars depict instances where the performance measure between the physicians and model are significantly different. The TN, Recall/TP rate, and Precision for the two cohorts combined is shown in the Appendix in Figure 6.

cancer. We tested three variations of the POMDP, considering observations as being independent over time (i.e., an exhaustive search of every combination of observations), as probabilities of a static regression model, and as probabilities derived from a DBN. Representing these variables dynamically via the DBN improved model performance in comparison with the logistic regression model and performed similarly as exhaustive search. This analysis is presented in the Appendix in Table XVI). Modeling observations using a dynamic model has two main advantages: first, a dynamic model can capture changes over time in these features, which in our opinion are potential indicators of lung cancer; and second, it allows effective scaling of the observation space with the incorporation of multiple temporal inputs. While considering temporal change is intuitive, many lung cancer risk models are "static" and use only the most current observations when calculating the likelihood of disease. Still, such risk models are useful in baseline assessment. The initial belief for each case in our POMDP uses the Tammemägi model [19], instantiated using the subject's own demographic and clinical variables at

baseline, updated with subsequent imaging observations.

The POMDP we designed makes use of a reward function learned through analysis of physicians' past decisions. We recently presented an adaptive maximum entropy inverse reinforcement learning (MaxEnt IRL) algorithm to inform a reward function in different cancers [12]. Using MaxEnt IRL, we established an optimization function explicitly modeling experts' actions. This strategy is different from other health-related POMDP applications [8], [10], [11] that typically employ cost functions based on quality-adjusted life years (QALYs), resource utilization, or other abstract metrics reflecting broader policy considerations. Building atop experts' prior actions, we take advantage of their experience and insights to integrate and weigh disparate information about a given individual; and by learning from multiple physicians and patients, we overcome potential biases. Yet curiously, per the diverging kappa score analysis, the POMDP is not fully replicating physicians' decisions. When it comes to early cancer prediction (e.g., predicting screening t_2 cancer from screening t_0), the POMDP outperforms experts, indicating that the model and associated reward function are discriminating between positive and negative cases in a different way. This difference may be attributed to the dynamic observation model used with this POMDP; when independent observations are instead assumed, we have found kappa scores to 1 in other domains, indicating high correlation between the model and experts' decisions [12]. Indeed, error analysis of the POMDP's FPs shows a different subset from the physicians: cases with smaller nodule sizes but more years of smoking and older baseline age are predicted as false positives by the POMDP. Early true positive cases share the same distributions, however, suggesting that a portion of POMDP false positives are early true positives. Table II illustrates this point in screening t_0 and post-screening for action a_0: 71% of TPs are being predicted from a_0 for post-screening cases – but if compared with screening t_0 cancer cases, they would have been classified as FPs.

Our previous work on predicting lung cancer in the LDCT screening setting showed encouraging results with earlier detection [16]. We showed that using a DBN trained on the NLST dataset we can match physicians' performance in predicting lung cancer, and in a portion of cases, in advance of the expert. But that method suffered from two limitations: first, the need to set an acceptable threshold for predicting lung cancer; and second, a decision-making process based solely on immediate outcomes without regard for longer-term benefits to the patient. We compared our current POMDP with our DBN [16], reproducing it on the same cohort of subjects used in this paper (i.e., using identical training and test sets and the same stratified five-fold cross-validation analysis). Even when setting different probability thresholds to generate performance metrics ($7 \cdot 10^{-6}$, 0.01, and 0.01 for each screening time point of the NLST study), our new POMDP-based approach outperformed the earlier model in terms of reducing the FP rate and improving early lung cancer prediction (see Table XVI in Appendix).

Limitations of this work are around the real-world nature of cancer surveillance. It is unlikely that patients are screened at fixed one-year time intervals, for any number of reasons. As such, a discrete time model may not be well-suited for instances of imaging observations at irregular frequencies. Alternatively, a continuous time model may address this issue more accurately. We also used a simplified, expert-defined three-state cancer state space (e.g., no cancer, uncertain cancer, lung cancer); a more sophisticated approach would involve learning this state space from the data, which we plan to explore in the future. Likewise, the observation space of our POMDP model is discrete, whereas a continuous value space might yield further improvements. This method can be explored through the use of linear Gaussian conditional probability tables (CPTs) instead of discrete observational CPTs. Lastly, the number of cancer and non-cancer cases changes as a function of time (i.e., more cancer cases are found at baseline). We did not account for this imbalance during training other than performing a stratified five-fold cross-validation to obtain an unbiased estimate of the model. Similarly, other temporal studies have used a k-fold cross validation to assess model performance [32]–[36]. This data imbalance over time occurred as a result of simplifying our lung POMDP model to consider only cases reporting a single pulmonary nodule over the course of the trial. A more concrete analysis would include cases with multiple nodules over time. However, it was not possible to ascertain the history of individual nodules in patients with multiple nodules as the NLST dataset does not contain sufficient tracking information on nodules. Moreover, the imputation of observations by our DBN observational model at six month intervals, even though it reduces over-screening, is inferred rather than based on true screening observations.

Future work includes conducting an external validation study of this NLST-based POMDP using data curated from our institution, expanding our observational model to consider multiple IPNs, as well as incorporating a richer set of imaging features derived from deep learning, which have demonstrated high classification performance in detecting malignant pulmonary nodules [37], [38].

REFERENCES

[1] American Cancer Society, "Cancer Facts and Figures 2017," 2017.

[2] National Lung Screening Trial Research Team, D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, J. D. Sicks, and P. P.-t.-p. Transmission, "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening," The New England journal of medicine, vol. 365, pp. 333–340, 8 2011.

[3] B. Bunn, "IASLC Issues Statement on Lung Cancer Screening with Low-Dose Computed Tomography."

[4] Z. Saghir, A. Dirksen, H. Ashraf, K. S. Bach, J. Brodersen, P. F. Clementsen, M. Døssing, H. Hansen, K. F. Kofoed, K. R. Larsen, J. Mortensen, J. F. Rasmussen, N. Seersholm, B. G. Skov, H. Thorsen, P. Tønnesen, and J. H. Pedersen, "CT screening for lung cancer brings forward early disease. The randomised Danish lung cancer screening trial: Status after five annual screening rounds with low-dose CT," Thorax, vol. 67, pp. 296–301, 4 2012.

[5] D. R. Aberle, S. DeMello, C. D. Berg, W. C. Black, B. Brewer, T. R. Church, K. L. Clingan, F. Duan, R. M. Fagerstrom, I. F. Gareen, C. A. Gatsonis, D. S. Gierada, A. Jain, G. C. Jones, I. Mahon, P. M. Marcus, J. M. Rathmell, J. Sicks, D. Sarah, C. D. Berg, W. C. Black, B. Brewer, T. R. Church, K. L. Clingan, F. Duan, R. M. Fagerstrom, I. F. Gareen, C. A. Gatsonis, D. S. Gierada, A. Jain, G. C. Jones, I. Mahon, P. M. Marcus, J. M. Rathmell, J. Sicks, and N. Team, "Results of the two

incidence screenings in the National Lung Screening Trial," N. Engl. J. Med., vol. 369, no. 10, pp. 920–931, 2013.

[6] J. Brodersen, L. M. Schwartz, C. Heneghan, J. W. O'Sullivan, J. K. Aronson, and S. Woloshin, "Overdiagnosis: what it is and what it isn't," 2018.

[7] M. Wiering and M. van Otterlo, Reinforcement Learning, vol. 12. 2012.

[8] M. Leshno, Z. Halpern, and N. Arber, "Cost-effectiveness of colorectal cancer screening in the average risk population," Health care management science, vol. 6, no. 3, pp. 165–174, 2003.

[9] T. Ayer, O. Alagoz, and N. K. Stout, "OR Forum—A POMDP Approach to Personalize Mammography Screening Decisions," Operations Research, vol. 60, no. 5, pp. 1019–1034, 2012.

[10] F. S. Erenay, O. Alagoz, and A. Said, "Optimizing Colonoscopy Screening for Colorectal Cancer Prevention and Surveillance," Manufacturing & Service Operations Management, vol. 16, no. 3, pp. 381–400, 2014.

[11] J. Zhang, B. T. Denton, H. Balasubramanian, N. D. Shah, and B. A. Inman, "Optimization of Prostate Biopsy Referral Decisions," Manufacturing & Service Operations Management, vol. 14, no. 4, pp. 529–547, 2012.

[12] P. Petousis, S. X. Han, W. Hsu, and A. A. Bui, "Generating Reward Functions using IRL Towards Individualized Cancer Screening," in CEUR Workshop Proceedings, vol. 2142, pp. 109–120, 2018.

[13] P. P. Massion and R. C. Walker, "Indeterminate pulmonary nodules: Risk for having or for developing lung cancer?," 12 2014.

[14] M. C. Tammemägi, H. A. Katki, W. G. Hocking, T. R. Church, N. Caporaso, P. A. Kvale, A. K. Chaturvedi, G. A. Silvestri, T. L. Riley, J. Commins, and C. D. Berg, "Selection Criteria for Lung-Cancer Screening," New England Journal of Medicine, vol. 368, no. 8, pp. 728–736, 2013.

[15] P. Petousis, A. Naeim, A. Mosleh, and W. Hsu, "Evaluating the Impact of Uncertainty on Risk Prediction: Towards More Robust Prediction Models," in AMIA - Annual Symposium proceedings, 2018.

[16] P. Petousis, S. X. Han, D. Aberle, and A. A. Bui, "Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: A dynamic Bayesian network," Artificial Intelligence in Medicine, vol. 72, pp. 42–55, 2016.

[17] K. P. Murphy, "The Bayes Net Toolbox for Matlab," Computing Science and Statistics, vol. 33, no. 2, p. 1024–1034, 2001.

[18] D. Koller and N. Friedman, Probabilistic Graphical Models - Principles and Techniques, vol. 53. 1989.

[19] M. C. Tammemagi, S. C. Lam, A. M. Mcwilliams, and D. D. Sin, "Incremental Value of Pulmonary Function and Sputum DNA Image Cytometry in Lung Cancer Risk Prediction," vol. 4, no. April, pp. 552–562, 2011.

[20] S. Russell and P. Norvig, Artificial Intelligence A Modern Approach (3rd Edition). 2010.

[21] A. R. Cassandra, M. L. Littman, and N. L. Zhang, "Incremental Pruning: A Simple, Fast, Exact Method for Partially Observable Markov Decision Processes," Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, pp. 54–61, 2 1997.

[22] E. J. Sondik and N. M. Apr, "The Optimal Control of Partially Observable Markov Processes Over the Infinite Horizon : Discounted Costs The Optimal Control of Partially Observable Markov," Source: Operations Research Operations Research Society of America, vol. 26, no. 2, pp. 282–304, 1978.

[23] C. C. White, "A survey of solution techniques for the partially observed Markov decision process," Annals of Operations Research, vol. 32, no. 1, pp. 215–230, 1991.

[24] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, vol. 136. LAWRENCE ERLBAUM ASSOCIATES, second edi ed., 1988.

[25] M. Tomczak and E. Tomczak, "The need to report effect size estimates revisited. An overview of some recommended measures of effect size," Trends in Sport Sciences, vol. 1, no. 21, pp. 19–25, 2014.

[26] C. O. Fritz, P. E. Morris, and J. J. Richler, "Effect size estimates: Current use, calculations, and interpretation," Journal of Experimental Psychology: General, vol. 141, pp. 2–18, 2 2012.

[27] C. S. White, E. Dharaiya, E. Campbell, and L. Boroczky, "The Vancouver Lung Cancer Risk Prediction Model: Assessment by Using a Subset of the National Lung Screening Trial Cohort," Radiology, vol. 283, p. 152627, 2016.

[28] B. J. McKee, S. M. Regis, A. B. McKee, S. Flacke, and C. Wald, "Performance of ACR Lung-RADS in a Clinical CT Lung Screening Program," Journal of the American College of Radiology, vol. 13, pp. R25–R29, 3 2016.

[29] P. F. Pinsky, D. S. Gierada, W. Black, R. Munden, H. Nath, D. Aberle, and E. Kazerooni, "Performance of lung-RADS in the national lung screening trial: A retrospective assessment," Annals of Internal Medicine, vol. 162, no. 7, pp. 485–491, 2015.

[30] C. I. Henschke, R. Yip, D. F. Yankelevitz, and J. P. Smith, "Definition of a positive test result in computed tomography screening for lung cancer," Annals of Internal Medicine, vol. 158, no. 4, pp. 246–252, 2013.

[31] American Lung Association, "Lung Cancer Fact Sheet," 2018.

[32] A. M. Alaa, K. H. Moon, W. Hsu, and M. Van Der Schaar, "ConfidentCare: A Clinical Decision Support System for Personalized Breast Cancer Screening," IEEE Transactions on Multimedia, vol. 18, no. 10, pp. 1942–1955, 2016.

[33] E. S. Burnside, J. Davis, J. Chhatwal, O. Alagoz, M. J. Lindstrom, B. M. Geller, B. Littenberg, K. a. Shaffer, C. E. Kahn, and C. D. Page, "Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings.," Radiology, vol. 251, no. 3, pp. 663–672, 2009.

[34] G. Cuaya, A. Mu??oz-Mel??ndez, L. N. Carrera, E. F. Morales, I. Qui??ones, A. I. P??rez, and A. Alessi, "A dynamic Bayesian network for estimating the risk of falls from real gait data," Medical and Biological Engineering and Computing, vol. 51, no. 1-2, pp. 29–37, 2013.

[35] E. W. Watt and A. A. T. Bui, "Evaluation of a dynamic bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative.," in AMIA 2008 Symposium, pp. 788–92, 2008.

[36] M. Van der Heijden, M. Velikova, and P. J. F. Lucas, "Learning Bayesian networks for clinical time series analysis," Journal of Biomedical Informatics, vol. 48, pp. 94–105, 2014.

[37] S. Shen, S. X. Han, D. R. Aberle, A. A. T. Bui, and W. Hsu, "An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification," arXiv preprint arXiv:1806.00712, 6 2018.

[38] A. Murphy, M. Skalski, and F. Gaillard, "The utilisation of convolutional neural networks in detecting pulmonary nodules: a review," The British Journal of Radiology, vol. 91, p. 20180028, 10 2018.

APPENDIX

ALGORITHM PSEUDO-CODES:

Algorithm 1: QMDP Algorithm

Input: S, A, R, ϵ
Output: Q matrix
Computing the Q matrix;
$V(s) = MDP\_VI(S, A, R, \epsilon)$;
for $s_i \in S$ **do**
  for $a \in A$ **do**
    $Q(s_i, a) = R(s, a) + \sum_{s_j \in S} T(s_j, a, s_i)V(s_i)$;
  end
end
return Q

Algorithm 2: Action selection Algorithm

Input: Q, b
Output: a_{opt} optimal action
Given belief b;
$a_{opt} = argmax_a \sum_{s_i \in S} b(s_i)Q(s_i, a)$
return a_{opt}

EFFECT SIZE RANGE

Cramer's V was used for χ^2 and Fisher tests.
r^2 or η^2 was used for Student test.
Cohen's r^2 was used for Wilcoxon-Mann-Whitney test.

TABLE IV: Magnitude of effect size, Cohen et al [24]. Tables 5-14 follow the color coding and bolding depicted in Table 4.

Magnitude of effect size	Cramer's V or ϕ	Cohen's d	r^2 or η^2
Small	0.1	0.2	0.01
Medium	0.3	0.5	0.059
Large	0.5	0.8	0.14

	Physician	POMDP	p	effect-size
cigsmok			$<$ **0.001**	0.432
0	527/735 (71.7 %)	237/839 (28.25 %)		
1	208/735 (28.3 %)	602/839 (71.75 %)		
diagcopd			$<$ **0.001**	0.187
0	728/733 (99.32 %)	758/834 (90.89 %)		
1	5/733 (0.68 %)	76/834 (9.11 %)		
famHist			$<$ **0.001**	0.327
0	670/735 (91.16 %)	530/839 (63.17 %)		
1	65/735 (8.84 %)	309/839 (36.83 %)		
pCancHist			$<$ **0.001**	0.152
0	728/735 (99.05 %)	778/839 (92.73 %)		
1	7/735 (0.95 %)	61/839 (7.27 %)		
gender			0.185	0.033
1	435/735 (59.18 %)	525/839 (62.57 %)		
2	300/735 (40.82 %)	314/839 (37.43 %)		
race			$<$ **0.001**	0.09
1	701/735 (95.37 %)	746/839 (88.92 %)		
2	11/735 (1.5 %)	75/839 (8.94 %)		
4	20/735 (2.72 %)	10/839 (1.19 %)		
5	0/735 (0 %)	5/839 (0.6 %)		
6	3/735 (0.41 %)	3/839 (0.36 %)		
educat			$<$ **0.001**	0.129
1	1/577 (0.17 %)	19/617 (3.08 %)		
2	9/577 (1.56 %)	79/617 (12.8 %)		
4	84/577 (14.56 %)	120/617 (19.45 %)		
5	180/577 (31.2 %)	211/617 (34.2 %)		
6	140/577 (24.26 %)	104/617 (16.86 %)		
7	152/577 (26.34 %)	67/617 (10.86 %)		
8	11/577 (1.91 %)	17/617 (2.76 %)		
sctpreatt0			$<$ **0.001**	0.078
1	566/664 (85.24 %)	213/280 (76.07 %)		
2	57/664 (8.58 %)	50/280 (17.86 %)		
3	35/664 (5.27 %)	15/280 (5.36 %)		
4	6/664 (0.9 %)	2/280 (0.71 %)		
sctmargins0			$<$ **0.001**	0.239
1	1/684 (0.15 %)	38/276 (13.77 %)		
2	603/684 (88.16 %)	184/276 (66.67 %)		
3	80/684 (11.7 %)	54/276 (19.57 %)		
BMI	29.06(4.94), 2	26.4(4.26), 3	$<$ **0.001**	0.072
smokeIntensity	29.16(11.57), 0	28.58(11.35), 0	0.24	$<$ 0.001
smokeyr	34.65(5.07), 14	45.77(6.38), 1	$<$ **0.001**	**0.514**
smokeQuitTime	5.72(5.36), 27	1.76(3.79), 9	$<$ **0.001**	**0.186**
age	58.39(3.27), 57	64.57(5.29), 49	$<$ **0.001**	**0.321**
LargestDiam0	5.39(1.88), 0	2.71(4.85), 0	$<$ **0.001**	**0.215**

TABLE V: Comparison between physicians and POMDP (baseline screen). For quantitative covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage)". Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

FALSE POSITIVES ANALYSIS

Testing data.

	Physician	POMDP	p	effect-size
cigsmok			$<$ **0.001**	0.343
0	605/875 (69.14 %)	279/802 (34.79 %)		
1	270/875 (30.86 %)	523/802 (65.21 %)		
diagcopd			$<$ **0.001**	0.137
0	862/872 (98.85 %)	745/796 (93.59 %)		
1	10/872 (1.15 %)	51/796 (6.41 %)		
famHist			$<$ **0.001**	0.259
0	774/875 (88.46 %)	537/802 (66.96 %)		
1	101/875 (11.54 %)	265/802 (33.04 %)		
pCancHist			$<$ **0.001**	0.107
0	862/875 (98.51 %)	758/802 (94.51 %)		
1	13/875 (1.49 %)	44/802 (5.49 %)		
gender			**0.037**	0.051
1	504/875 (57.6 %)	503/802 (62.72 %)		
2	371/875 (42.4 %)	299/802 (37.28 %)		
race			$<$ **0.001**	0.074
1	826/875 (94.4 %)	719/802 (89.65 %)		
2	21/875 (2.4 %)	68/802 (8.48 %)		
4	25/875 (2.86 %)	10/802 (1.25 %)		
5	0/875 (0 %)	1/802 (0.12 %)		
6	3/875 (0.34 %)	4/802 (0.5 %)		
educat			$<$ **0.001**	0.099
1	2/685 (0.29 %)	14/599 (2.34 %)		
2	11/685 (1.61 %)	61/599 (10.18 %)		
4	101/685 (14.74 %)	105/599 (17.53 %)		
5	217/685 (31.68 %)	199/599 (33.22 %)		
6	166/685 (24.23 %)	123/599 (20.53 %)		
7	173/685 (25.26 %)	85/599 (14.19 %)		
8	15/685 (2.19 %)	12/599 (2 %)		
sctpreatt0			$<$ **0.001**	0.087
1	557/651 (85.56 %)	181/244 (74.18 %)		
2	58/651 (8.91 %)	48/244 (19.67 %)		
3	32/651 (4.92 %)	13/244 (5.33 %)		
4	4/651 (0.61 %)	2/244 (0.82 %)		
sctpreatt1			**0.023**	0.056
1	528/595 (88.74 %)	339/406 (83.5 %)		
2	47/595 (7.9 %)	56/406 (13.79 %)		
3	16/595 (2.69 %)	9/406 (2.22 %)		
4	4/595 (0.67 %)	2/406 (0.49 %)		
sctmargins0			$<$ **0.001**	0.233
1	0/675 (0 %)	30/240 (12.5 %)		
2	595/675 (88.15 %)	164/240 (68.33 %)		
3	80/675 (11.85 %)	46/240 (19.17 %)		
sctmargins1			$<$ **0.001**	0.145
1	7/620 (1.13 %)	38/416 (9.13 %)		
2	525/620 (84.68 %)	303/416 (72.84 %)		
3	88/620 (14.19 %)	75/416 (18.03 %)		
BMI	28.9(5.06), 4	26.9(4.63), 3	$<$ **0.001**	0.041
smokeIntensity	28.77(11.26), 0	28.63(11.47), 0	0.545	$<$ 0.001
smokeyr	35.24(5.29), 17	43.51(7.13), 2	$<$ **0.001**	**0.323**
smokeQuitTime	5.4(5.36), 32	2.29(4.28), 11	$<$ **0.001**	0.112
age	58.62(3.49), 69	63.05(5.53), 56	$<$ **0.001**	**0.166**
LargestDiam0	4.36(2.55), 0	2.45(4.7), 0	$<$ **0.001**	0.133
LargestDiam1	4(3.59), 31	4.1(5.56), 38	0.297	$<$ 0.001

TABLE VI: Comparison between physicians and POMDP (2^{nd} screen). For quantitative covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage)". Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

EARLY TPS ANALYSIS

Testing data.

	Physician	POMDP	p	effect-size
cigsmok			$<$ **0.001**	0.278
0	643/948 (67.83 %)	306/766 (39.95 %)		
1	305/948 (32.17 %)	460/766 (60.05 %)		
diagcopd			$<$ **0.001**	0.074
0	928/946 (98.1 %)	728/763 (95.41 %)		
1	18/946 (1.9 %)	35/763 (4.59 %)		
famHist			$<$ **0.001**	0.191
0	830/948 (87.55 %)	554/766 (72.32 %)		
1	118/948 (12.45 %)	212/766 (27.68 %)		
pCancHist			$<$ **0.001**	0.08
0	931/948 (98.21 %)	730/766 (95.3 %)		
1	17/948 (1.79 %)	36/766 (4.7 %)		
gender			0.079	0.042
1	547/948 (57.7 %)	475/766 (62.01 %)		
2	401/948 (42.3 %)	291/766 (37.99 %)		
race			$<$ **0.001**	0.071
1	894/948 (94.3 %)	690/766 (90.08 %)		
2	26/948 (2.74 %)	61/766 (7.96 %)		
4	25/948 (2.64 %)	12/766 (1.57 %)		
6	3/948 (0.32 %)	3/766 (0.39 %)		
educat			$<$ **0.001**	0.085
1	1/746 (0.13 %)	13/567 (2.29 %)		
2	11/746 (1.47 %)	43/567 (7.58 %)		
4	116/746 (15.55 %)	99/567 (17.46 %)		
5	236/746 (31.64 %)	184/567 (32.45 %)		
6	184/746 (24.66 %)	127/567 (22.4 %)		
7	182/746 (24.4 %)	89/567 (15.7 %)		
8	16/746 (2.14 %)	12/567 (2.12 %)		
sctpreatt0			$<$ **0.001**	0.084
1	543/627 (86.6 %)	177/235 (75.32 %)		
2	54/627 (8.61 %)	44/235 (18.72 %)		
3	26/627 (4.15 %)	12/235 (5.11 %)		
4	4/627 (0.64 %)	2/235 (0.85 %)		
sctpreatt1			$<$ **0.001**	0.066
1	528/588 (89.8 %)	301/361 (83.38 %)		
2	43/588 (7.31 %)	51/361 (14.13 %)		
3	13/588 (2.21 %)	8/361 (2.22 %)		
4	4/588 (0.68 %)	1/361 (0.28 %)		
sctpreatt2			**0.02**	0.055
1	503/572 (87.94 %)	403/490 (82.24 %)		
2	52/572 (9.09 %)	74/490 (15.1 %)		
3	11/572 (1.92 %)	10/490 (2.04 %)		
4	6/572 (1.05 %)	3/490 (0.61 %)		
sctmargins0			$<$ **0.001**	0.234
1	0/650 (0 %)	29/232 (12.5 %)		
2	578/650 (88.92 %)	160/232 (68.97 %)		
3	72/650 (11.08 %)	43/232 (18.53 %)		
sctmargins1			$<$ **0.001**	0.176
1	0/611 (0 %)	32/370 (8.65 %)		
2	529/611 (86.58 %)	272/370 (73.51 %)		
3	82/611 (13.42 %)	66/370 (17.84 %)		
sctmargins2			$<$ **0.001**	0.099
1	8/582 (1.37 %)	30/500 (6 %)		
2	501/582 (86.08 %)	389/500 (77.8 %)		
3	73/582 (12.54 %)	81/500 (16.2 %)		
BMI	28.83(5.05), 6	27.3(4.82), 2	$<$ **0.001**	0.025
smokeIntensity	28.81(11.19), 0	29.15(11.94), 0	0.889	$<$ 0.001
smokeyr	35.51(5.46), 18	41.62(7.25), 4	$<$ **0.001**	**0.199**
smokeQuitTime	5.24(5.39), 32	2.82(4.74), 16	$<$ **0.001**	0.068
age	58.69(3.58), 77	61.81(5.33), 57	$<$ **0.001**	0.088
LargestDiam0	3.85(2.75), 0	2.46(4.72), 0	$<$ **0.001**	0.083
LargestDiam1	3.61(3.61), 30	3.79(5.43), 38	0.442	$<$ 0.001
LargestDiam2	3.34(2.58), 54	4.74(4.17), 59	$<$ **0.001**	0.044

TABLE VII: Comparison between physicians and POMDP (3^{rd} screen). For quantitative covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage)". Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

	Physician	POMDP	p	effect-size
cigsmok			$<$ **0.001**	0.276
0	639/947 (67.48 %)	299/753 (39.71 %)		
1	308/947 (32.52 %)	454/753 (60.29 %)		
diagcopd			$<$ **0.001**	0.072
0	927/945 (98.1 %)	716/750 (95.47 %)		
1	18/945 (1.9 %)	34/750 (4.53 %)		
famHist			$<$ **0.001**	0.193
0	830/947 (87.65 %)	544/753 (72.24 %)		
1	117/947 (12.35 %)	209/753 (27.76 %)		
pCancHist			$<$ **0.001**	0.079
0	930/947 (98.2 %)	718/753 (95.35 %)		
1	17/947 (1.8 %)	35/753 (4.65 %)		
gender			0.079	0.043
1	545/947 (57.55 %)	466/753 (61.89 %)		
2	402/947 (42.45 %)	287/753 (38.11 %)		
race			$<$ **0.001**	0.069
1	892/947 (94.19 %)	678/753 (90.04 %)		
2	27/947 (2.85 %)	60/753 (7.97 %)		
4	25/947 (2.64 %)	12/753 (1.59 %)		
6	3/947 (0.32 %)	3/753 (0.4 %)		
educat			$<$ **0.001**	0.082
1	2/745 (0.27 %)	12/556 (2.16 %)		
2	11/745 (1.48 %)	42/556 (7.55 %)		
4	116/745 (15.57 %)	96/556 (17.27 %)		
5	233/745 (31.28 %)	183/556 (32.91 %)		
6	184/745 (24.7 %)	125/556 (22.48 %)		
7	183/745 (24.56 %)	88/556 (15.83 %)		
8	16/745 (2.15 %)	10/556 (1.8 %)		
sctpreatt0			$<$ **0.001**	0.086
1	544/628 (86.62 %)	174/232 (75 %)		
2	54/628 (8.6 %)	44/232 (18.97 %)		
3	26/628 (4.14 %)	12/232 (5.17 %)		
4	4/628 (0.64 %)	2/232 (0.86 %)		
sctpreatt1			$<$ **0.001**	0.065
1	528/588 (89.8 %)	298/356 (83.71 %)		
2	43/588 (7.31 %)	50/356 (14.04 %)		
3	13/588 (2.21 %)	7/356 (1.97 %)		
4	4/588 (0.68 %)	1/356 (0.28 %)		
sctpreatt2			**0.019**	0.056
1	503/571 (88.09 %)	396/481 (82.33 %)		
2	51/571 (8.93 %)	72/481 (14.97 %)		
3	11/571 (1.93 %)	10/481 (2.08 %)		
4	6/571 (1.05 %)	3/481 (0.62 %)		
sctmargins0			$<$ **0.001**	0.229
1	0/651 (0 %)	27/229 (11.79 %)		
2	579/651 (88.94 %)	159/229 (69.43 %)		
3	72/651 (11.06 %)	43/229 (18.78 %)		
sctmargins1			$<$ **0.001**	0.174
1	0/611 (0 %)	31/365 (8.49 %)		
2	529/611 (86.58 %)	270/365 (73.97 %)		
3	82/611 (13.42 %)	64/365 (17.53 %)		
sctmargins2			$<$ **0.001**	0.1
1	8/581 (1.38 %)	30/491 (6.11 %)		
2	501/581 (86.23 %)	383/491 (78 %)		
3	72/581 (12.39 %)	78/491 (15.89 %)		
BMI	28.82(5.06), 6	27.32(4.82), 2	$<$ **0.001**	0.024
smokeIntensity	28.77(11.19), 0	29.02(11.83), 0	0.961	$<$ 0.001
smokeyr	35.52(5.47), 18	41.64(7.25), 4	$<$ **0.001**	**0.199**
smokeQuitTime	5.21(5.38), 32	2.79(4.72), 16	$<$ **0.001**	0.068
age	58.68(3.56), 76	61.79(5.31), 56	$<$ **0.001**	0.088
LargestDiam0	3.86(2.74), 0	2.47(4.74), 0	$<$ **0.001**	0.083
LargestDiam1	3.61(3.61), 30	3.8(5.44), 37	0.442	$<$ 0.001
LargestDiam2	3.35(2.58), 55	4.73(4.17), 58	$<$ **0.001**	0.043

TABLE VIII: Comparison between physicians and POMDP (post screening). For quantitative covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage)". Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

	Physician	POMDP	p	effect-size
cigsmok			0.084	0.229
0	13/30 (43.33 %)	5/27 (18.52 %)		
1	17/30 (56.67 %)	22/27 (81.48 %)		
diagcopd			0.238	0.127
0	28/30 (93.33 %)	22/27 (81.48 %)		
1	2/30 (6.67 %)	5/27 (18.52 %)		
famHist			0.855	0.024
0	16/30 (53.33 %)	16/27 (59.26 %)		
1	14/30 (46.67 %)	11/27 (40.74 %)		
pCancHist			0.66	0.016
0	28/30 (93.33 %)	24/27 (88.89 %)		
1	2/30 (6.67 %)	3/27 (11.11 %)		
gender			0.098	0.219
1	16/30 (53.33 %)	21/27 (77.78 %)		
2	14/30 (46.67 %)	6/27 (22.22 %)		
race			0.238	0.127
1	28/30 (93.33 %)	22/27 (81.48 %)		
2	2/30 (6.67 %)	5/27 (18.52 %)		
educat			0.838	0.106
1	1/21 (4.76 %)	1/19 (5.26 %)		
2	1/21 (4.76 %)	3/19 (15.79 %)		
4	4/21 (19.05 %)	4/19 (21.05 %)		
5	7/21 (33.33 %)	7/19 (36.84 %)		
6	4/21 (19.05 %)	2/19 (10.53 %)		
7	4/21 (19.05 %)	2/19 (10.53 %)		
sctpreatt0			0.324	0.218
1	18/29 (62.07 %)	6/6 (100 %)		
2	6/29 (20.69 %)	0/6 (0 %)		
3	5/29 (17.24 %)	0/6 (0 %)		
sctmargins0			0.87	0.101
1	11/27 (40.74 %)	4/7 (57.14 %)		
2	7/27 (25.93 %)	1/7 (14.29 %)		
3	9/27 (33.33 %)	2/7 (28.57 %)		
BMI	27.15(7.99), 0	24.54(3.61), 0	0.198	0.029
smokeIntensity	30.17(11.56), 0	29.93(11.08), 0	0.98	$<$ 0.001
smokeyr	43.83(5.79), 0	45.7(6.01), 0	0.238	0.025
smokeQuitTime	1.86(3.1), 1	0.73(2.6), 1	0.059	0.065
age	63.83(4.73), 6	63.92(5.36), 2	0.952	$<$ 0.001
LargestDiam0	13.83(20.18), 0	3.52(6.67), 0	$<$ **0.001**	**0.319**

TABLE IX: Comparison between physicians and POMDP (early prediction of 2^{nd} screen with a_0). For quantitatives covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage). Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

	Physician	POMDP	p	effect-size
cigsmok			**0.016**	0.31
0	14/26 (53.85 %)	7/34 (20.59 %)		
1	12/26 (46.15 %)	27/34 (79.41 %)		
diagcopd			1	< 0.001
0	23/26 (88.46 %)	30/34 (88.24 %)		
1	3/26 (11.54 %)	4/34 (11.76 %)		
famHist			0.101	0.212
0	20/26 (76.92 %)	18/34 (52.94 %)		
1	6/26 (23.08 %)	16/34 (47.06 %)		
pCancHist			1	< 0.001
0	25/26 (96.15 %)	32/34 (94.12 %)		
1	1/26 (3.85 %)	2/34 (5.88 %)		
gender			0.725	0.045
1	14/26 (53.85 %)	21/34 (61.76 %)		
2	12/26 (46.15 %)	13/34 (38.24 %)		
race			1	0.094
1	25/26 (96.15 %)	31/34 (91.18 %)		
2	1/26 (3.85 %)	1/34 (2.94 %)		
4	0/26 (0 %)	1/34 (2.94 %)		
6	0/26 (0 %)	1/34 (2.94 %)		
educat			0.053	0.208
2	2/21 (9.52 %)	4/27 (14.81 %)		
4	6/21 (28.57 %)	7/27 (25.93 %)		
5	4/21 (19.05 %)	13/27 (48.15 %)		
6	4/21 (19.05 %)	1/27 (3.7 %)		
7	5/21 (23.81 %)	1/27 (3.7 %)		
8	0/21 (0 %)	1/27 (3.7 %)		
sctpreatt0			1	0.065
1	16/26 (61.54 %)	2/3 (66.67 %)		
2	8/26 (30.77 %)	1/3 (33.33 %)		
4	2/26 (7.69 %)	0/3 (0 %)		
sctmargins0			1	0.11
1	5/22 (22.73 %)	0/2 (0 %)		
2	9/22 (40.91 %)	1/2 (50 %)		
3	8/22 (36.36 %)	1/2 (50 %)		
BMI	$25.84(3.83), 0$	$25.4(5.28), 0$	0.438	0.01
smokeIntensity	$33.65(16.34), 0$	$29.12(12.34), 0$	0.346	0.015
smokeyr	$40.96(6.86), 0$	$47.59(7.57), 0$	$<$ **0.001**	**0.171**
smokeQuitTime	$4.12(5.32), 1$	$1.45(3.96), 1$	**0.01**	0.116
age	$61.32(4.5), 1$	$65.91(5.34), 2$	$<$ **0.001**	**0.167**
LargestDiam0	$10.5(7.63), 0$	$0.85(3.23), 0$	$<$ **0.001**	**0.71**

TABLE X: Comparison between physicians and POMDP (early prediction of 3^{rd} screen with a_0). For quantitatives covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage). Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

	Physician	POMDP	p	effect-size
cigsmok			0.051	0.226
0	20/43 (46.51 %)	7/32 (21.88 %)		
1	23/43 (53.49 %)	25/32 (78.12 %)		
diagcopd			0.451	0.048
0	40/43 (93.02 %)	28/32 (87.5 %)		
1	3/43 (6.98 %)	4/32 (12.5 %)		
famHist			0.258	0.131
0	32/43 (74.42 %)	19/32 (59.38 %)		
1	11/43 (25.58 %)	13/32 (40.62 %)		
pCancHist			0.572	0.03
0	42/43 (97.67 %)	30/32 (93.75 %)		
1	1/43 (2.33 %)	2/32 (6.25 %)		
gender			1	< 0.001
1	24/43 (55.81 %)	17/32 (53.12 %)		
2	19/43 (44.19 %)	15/32 (46.88 %)		
race			**0.038**	0.176
1	39/43 (90.7 %)	29/32 (90.62 %)		
2	4/43 (9.3 %)	0/32 (0 %)		
4	0/43 (0 %)	1/32 (3.12 %)		
6	0/43 (0 %)	2/32 (6.25 %)		
educat			0.195	0.156
2	3/33 (9.09 %)	4/25 (16 %)		
4	8/33 (24.24 %)	6/25 (24 %)		
5	9/33 (27.27 %)	12/25 (48 %)		
6	7/33 (21.21 %)	1/25 (4 %)		
7	5/33 (15.15 %)	1/25 (4 %)		
8	1/33 (3.03 %)	1/25 (4 %)		
sctpreatt0			1	0.065
1	16/26 (61.54 %)	2/3 (66.67 %)		
2	8/26 (30.77 %)	1/3 (33.33 %)		
4	2/26 (7.69 %)	0/3 (0 %)		
sctpreatt1			0.754	0.107
1	22/34 (64.71 %)	4/6 (66.67 %)		
2	7/34 (20.59 %)	1/6 (16.67 %)		
3	2/34 (5.88 %)	1/6 (16.67 %)		
4	3/34 (8.82 %)	0/6 (0 %)		
sctmargins0			1	0.11
1	5/22 (22.73 %)	0/2 (0 %)		
2	9/22 (40.91 %)	1/2 (50 %)		
3	8/22 (36.36 %)	1/2 (50 %)		
sctmargins1			0.645	0.126
1	9/31 (29.03 %)	2/6 (33.33 %)		
2	12/31 (38.71 %)	1/6 (16.67 %)		
3	10/31 (32.26 %)	3/6 (50 %)		
BMI	$26.34(3.32), 0$	$25.29(5.44), 0$	0.097	0.037
smokeIntensity	$32.44(15.13), 0$	$26.5(10), 0$	0.079	0.041
smokeyr	$42.4(6.47), 0$	$46.25(7.99), 0$	**0.029**	0.064
smokeQuitTime	$3.26(4.79), 1$	$1.55(4.07), 1$	**0.032**	0.063
age	$61.69(4.79), 1$	$64.83(5.86), 2$	**0.019**	0.076
LargestDiam0	$6.35(7.85), 0$	$0.91(3.32), 0$	$<$ **0.001**	**0.25**
LargestDiam1	$7.86(8.04), 1$	$2.98(7.15), 0$	$<$ **0.001**	**0.225**

TABLE XI: Comparison between physicians and POMDP (early prediction of 3^{rd} screen with a_1). For quantitatives covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage). Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

	Physician	POMDP	p	effect-size
cigsmok			0.789	0.032
0	14/45 (31.11 %)	9/24 (37.5 %)		
1	31/45 (68.89 %)	15/24 (62.5 %)		
diagcopd			0.687	< 0.01
0	41/45 (91.11 %)	21/24 (87.5 %)		
1	4/45 (8.89 %)	3/24 (12.5 %)		
famHist			0.39	0.104
0	34/45 (75.56 %)	15/24 (62.5 %)		
1	11/45 (24.44 %)	9/24 (37.5 %)		
pCancHist			0.333	0.089
0	43/45 (95.56 %)	21/24 (87.5 %)		
1	2/45 (4.44 %)	3/24 (12.5 %)		
gender			1	< 0.001
1	30/45 (66.67 %)	16/24 (66.67 %)		
2	15/45 (33.33 %)	8/24 (33.33 %)		
race			**0.012**	0.253
1	43/45 (95.56 %)	20/24 (83.33 %)		
2	0/45 (0 %)	4/24 (16.67 %)		
4	2/45 (4.44 %)	0/24 (0 %)		
educat			0.072	0.216
1	0/26 (0 %)	1/15 (6.67 %)		
2	0/26 (0 %)	2/15 (13.33 %)		
4	10/26 (38.46 %)	3/15 (20 %)		
5	9/26 (34.62 %)	5/15 (33.33 %)		
6	4/26 (15.38 %)	2/15 (13.33 %)		
7	3/26 (11.54 %)	0/15 (0 %)		
8	0/26 (0 %)	2/15 (13.33 %)		
sctpreatt0			0.731	0.155
1	28/41 (68.29 %)	5/5 (100 %)		
2	9/41 (21.95 %)	0/5 (0 %)		
3	4/41 (9.76 %)	0/5 (0 %)		
sctmargins0			0.287	0.175
1	11/43 (25.58 %)	3/5 (60 %)		
2	24/43 (55.81 %)	2/5 (40 %)		
3	8/43 (18.6 %)	0/5 (0 %)		
BMI	26.52(4.71), 0	27.08(4.83), 0	0.668	< 0.001
smokeIntensity	28.22(10.07), 0	31.12(13.3), 0	0.442	< 0.001
smokeyr	46.33(5.51), 0	45.92(7.26), 0	0.807	< 0.001
smokeQuitTime	2(4.13), 1	3.33(5.11), 0	0.387	0.011
age	65.36(5.32), 3	66.79(4.45), 0	0.246	0.021
LargestDiam0	8.02(4.91), 0	1.46(2.93), 0	$<$ **0.001**	**0.397**

TABLE XII: Comparison between physicians and POMDP (early prediction of post-screening with a_0). For quantitatives covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage). Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

	Physician	POMDP	p	effect-size
cigsmok			0.267	0.124
0	22/65 (33.85 %)	8/15 (53.33 %)		
1	43/65 (66.15 %)	7/15 (46.67 %)		
diagcopd			1	0.019
0	57/65 (87.69 %)	14/15 (93.33 %)		
1	8/65 (12.31 %)	1/15 (6.67 %)		
famHist			0.166	0.155
0	49/65 (75.38 %)	8/15 (53.33 %)		
1	16/65 (24.62 %)	7/15 (46.67 %)		
pCancHist			0.234	0.074
0	62/65 (95.38 %)	13/15 (86.67 %)		
1	3/65 (4.62 %)	2/15 (13.33 %)		
gender			1	< 0.001
1	43/65 (66.15 %)	10/15 (66.67 %)		
2	22/65 (33.85 %)	5/15 (33.33 %)		
race			**0.035**	0.241
1	61/65 (93.85 %)	12/15 (80 %)		
2	1/65 (1.54 %)	3/15 (20 %)		
4	3/65 (4.62 %)	0/15 (0 %)		
educat			0.361	0.148
1	0/35 (0 %)	1/12 (8.33 %)		
2	2/35 (5.71 %)	2/12 (16.67 %)		
4	11/35 (31.43 %)	4/12 (33.33 %)		
5	12/35 (34.29 %)	2/12 (16.67 %)		
6	6/35 (17.14 %)	1/12 (8.33 %)		
7	3/35 (8.57 %)	1/12 (8.33 %)		
8	1/35 (2.86 %)	1/12 (8.33 %)		
sctpreatt0			1	0.117
1	30/43 (69.77 %)	3/3 (100 %)		
2	9/43 (20.93 %)	0/3 (0 %)		
3	4/43 (9.3 %)	0/3 (0 %)		
sctpreatt1			0.28	0.108
1	37/48 (77.08 %)	4/7 (57.14 %)		
2	7/48 (14.58 %)	2/7 (28.57 %)		
3	4/48 (8.33 %)	1/7 (14.29 %)		
sctmargins0			0.268	0.156
1	12/45 (26.67 %)	2/3 (66.67 %)		
2	25/45 (55.56 %)	1/3 (33.33 %)		
3	8/45 (17.78 %)	0/3 (0 %)		
sctmargins1			0.745	0.092
1	8/47 (17.02 %)	1/7 (14.29 %)		
2	27/47 (57.45 %)	3/7 (42.86 %)		
3	12/47 (25.53 %)	3/7 (42.86 %)		
BMI	26.26(4.62), 0	27.13(5.05), 0	0.613	< 0.001
smokeIntensity	28.57(10.36), 0	32.33(12.66), 0	0.302	0.013
smokeyr	46.12(5.94), 0	42.6(6.29), 0	0.062	0.048
smokeQuitTime	2.31(4.29), 1	4.2(5.2), 0	0.116	0.031
age	65.53(5.41), 6	65(5.59), 1	0.763	< 0.001
LargestDiam0	5.78(5.44), 0	1.33(2.79), 0	$<$ **0.001**	0.128
LargestDiam1	6.53(4.74), 4	4.59(6.07), 1	0.111	0.034

TABLE XIII: Comparison between physicians and POMDP (early prediction of post-screening with a_1). For quantitatives covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage). Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

	Physician	POMDP	p	effect-size
cigsmok			0.309	0.107
0	27/78 (34.62 %)	7/13 (53.85 %)		
1	51/78 (65.38 %)	6/13 (46.15 %)		
diagcopd			1	< 0.001
0	70/78 (89.74 %)	12/13 (92.31 %)		
1	8/78 (10.26 %)	1/13 (7.69 %)		
famHist			1	< 0.001
0	58/78 (74.36 %)	10/13 (76.92 %)		
1	20/78 (25.64 %)	3/13 (23.08 %)		
pCancHist			1	< 0.001
0	73/78 (93.59 %)	12/13 (92.31 %)		
1	5/78 (6.41 %)	1/13 (7.69 %)		
gender			1	< 0.01
1	50/78 (64.1 %)	9/13 (69.23 %)		
2	28/78 (35.9 %)	4/13 (30.77 %)		
race			0.724	0.059
1	71/78 (91.03 %)	12/13 (92.31 %)		
2	4/78 (5.13 %)	1/13 (7.69 %)		
4	3/78 (3.85 %)	0/13 (0 %)		
educat			0.091	0.177
1	0/44 (0 %)	1/11 (9.09 %)		
2	3/44 (6.82 %)	1/11 (9.09 %)		
4	13/44 (29.55 %)	3/11 (27.27 %)		
5	16/44 (36.36 %)	1/11 (9.09 %)		
6	7/44 (15.91 %)	2/11 (18.18 %)		
7	4/44 (9.09 %)	1/11 (9.09 %)		
8	1/44 (2.27 %)	2/11 (18.18 %)		
sctpreatt0			1	0.117
1	30/43 (69.77 %)	3/3 (100 %)		
2	9/43 (20.93 %)	0/3 (0 %)		
3	4/43 (9.3 %)	0/3 (0 %)		
sctpreatt1			0.367	0.093
1	39/51 (76.47 %)	3/5 (60 %)		
2	8/51 (15.69 %)	1/5 (20 %)		
3	4/51 (7.84 %)	1/5 (20 %)		
sctpreatt2			1	0.055
1	32/42 (76.19 %)	7/9 (77.78 %)		
2	8/42 (19.05 %)	2/9 (22.22 %)		
3	1/42 (2.38 %)	0/9 (0 %)		
4	1/42 (2.38 %)	0/9 (0 %)		
sctmargins0			0.268	0.156
1	12/45 (26.67 %)	2/3 (66.67 %)		
2	25/45 (55.56 %)	1/3 (33.33 %)		
3	8/45 (17.78 %)	0/3 (0 %)		
sctmargins1			0.554	0.076
1	8/50 (16 %)	1/5 (20 %)		
2	29/50 (58 %)	2/5 (40 %)		
3	13/50 (26 %)	2/5 (40 %)		
sctmargins2			0.746	0.109
1	5/43 (11.63 %)	0/9 (0 %)		
2	27/43 (62.79 %)	6/9 (66.67 %)		
3	11/43 (25.58 %)	3/9 (33.33 %)		
BMI	$26.73(4.78), 0$	$26.52(5.19), 0$	0.755	< 0.001
smokeIntensity	$28.49(10.29), 0$	$36.54(15.99), 0$	0.075	0.035
smokeyr	$45.33(5.93), 0$	$40.31(7.78), 0$	**0.043**	0.053
smokeQuitTime	$2.29(4.24), 1$	$4.77(5.78), 0$	0.086	0.033
age	$64.89(5.46), 6$	$62.92(6.49), 1$	0.3	0.013
LargestDiam0	$4.82(5.41), 0$	$1.54(2.96), 0$	**0.027**	0.053
LargestDiam1	$5.79(5.14), 4$	$3.17(4.34), 1$	0.096	0.032
LargestDiam2	$5.24(5.33), 11$	$5.17(3.81), 1$	0.774	< 0.001

TABLE XIV: Comparison between physicians and POMDP (early prediction of post-screening with a_2). For quantitatives covariates: "mean (sd), missing data", and for categorical covariates: "effective/ total effective (percentage). Student test or Wilcoxon-Mann-Whitney test, χ^2 test or Fisher test used when appropriate. The effect-size column follows the color coding and bolding depicted in Table 4.

Variable	Variable Explanation	Categories
cigsmok	Smoking status at T0 Participant	0="Former" 1="Current"
diagcopd	COPD: Ever diagnosed prior to trial?	0="No" 1="Yes"
famHist	Family History of lung cancer, 1st degree relative	0="No" 1="Yes"
pCancHist	Personal cancer history, all types of cancer	0="No" 1="Yes"
gender		1="Male" 2="Female"
race		1="White" 2="Black" 3="Hispanic" 4="Asian" 5="American Indian or Alaskan Native" 6="Native Hawaiian or Other Pacific Islander"
educat	Education level	1="8th grade or less" 2="9th-11th grade" 3="High school graduate/GED" 4="Post high school training,excluding college" 5="Associate degree/ some college" 6="Bachelors Degree" 7="Graduate School" 8="Other"
BMI	Body mass Index	continuous
smokeIntensity	Average number of cigarettes per day	continuous
smokeYr	Total years of smoking	continuous
smokeQuitTime	Time of Quitting Smoking	continuous
age	Age at T0	continuous
sct_pre_att0_2	Predominant attenuation T0-2	1="Soft Tissue" 2="Ground Glass" 3="Mixed" 4="Other"
sct_margins0_2	Margins T0-2	1="Spiculated" 2="Smooth" 3="Poorly defined"
LargestDiam0_2	Largest nodule diameter (mm) T0-2	continuous

TABLE XV: The categories and description of each variable.

VARIABLES' CATEGORIES

	POMDP			**DBN 2016**		
	TN rate	**TP rate/Recall**	**Precision**	**TN rate**	**TP rate/Recall**	**Precision**
	Screening T0 (Cancers = 32, Non-Cancers = 1,047)					
a_0	0.47	0.97	0.05	0.43	1.00	0.06
	Screening T1 (Cancers = 17, Non-Cancers = 1,030)					
a_0	0.47	0.67	0.02	0.43	0.47	0.01
$a_{0.5}$	0.47	0.67	0.02			
a_1	0.34	0.98	0.02	0.30	0.99	0.02
	Screening T2 (Cancers = 21, Non-Cancers = 1,009)					
a_0	0.47	0.55	0.02	0.42	0.30	0.01
$a_{0.5}$	0.47	0.55	0.02			
a_1	0.34	0.69	0.02	0.30	0.48	0.01
$a_{1.5}$	0.34	0.70	0.02			
a_2	0.25	0.96	0.03	0.18	1.00	0.03
	Post Screening (Cancers = 19, Non-Cancers = 990)					
a_0	0.48	0.68	0.02	0.43	0.49	0.02
$a_{0.5}$	0.47	0.68	0.02			
a_1	0.35	0.81	0.02	0.30	0.71	0.02
$a_{1.5}$	0.34	0.81	0.02			
a_2	0.25	0.93	0.02	0.18	0.88	0.02

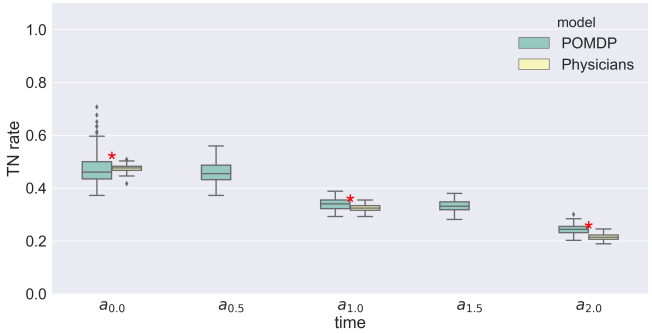
TABLE XVI: POMDP Vs DBN 2016 model. Testing data.

	POMDP – DBN			**POMDP – Exhaustive search**			**POMDP – LR**		
	TN rate	**TP rate/Recall**	**Precision**	**TN rate**	**TP rate/Recall**	**Precision**	**TN rate**	**TP rate/Recall**	**Precision**
Screening T0 (Cancers = 32, Non-Cancers = 1,047)									
a_0	0.84	0.83	0.14	0.76	0.87	0.1	0.9	0.79	0.2
Screening T1 (Cancers = 17, Non-Cancers = 1,030)									
a_0	0.84	0.31	0.03	0.76	0.36	0.03	0.9	0.21	0.03
$a_{0.5}$	0.84	0.31	0.03	0.7	0.74	0.04	0.88	0.45	0.06
a_1	0.74	0.69	0.04	0.7	0.8	0.04	0.85	0.63	0.07
Screening T2 (Cancers = 21, Non-Cancers = 1,009)									
a_0	0.84	0.19	0.02	0.77	0.23	0.02	0.9	0.12	0.02
$a_{0.5}$	0.84	0.19	0.02	0.7	0.38	0.03	0.88	0.15	0.03
a_1	0.75	0.34	0.03	0.69	0.4	0.03	0.85	0.22	0.03
$a_{1.5}$	0.74	0.36	0.03	0.64	0.74	0.04	0.84	0.37	0.05
a_2	0.69	0.71	0.05	0.63	0.8	0.04	0.84	0.37	0.05
Post Screening (Cancers = 19, Non-Cancers = 990)									
a_0	0.84	0.24	0.03	0.77	0.32	0.03	0.9	0.15	0.03
$a_{0.5}$	0.84	0.24	0.03	0.7	0.4	0.03	0.89	0.19	0.03
a_1	0.75	0.39	0.03	0.63	0.42	0.03	0.86	0.25	0.03
$a_{1.5}$	0.74	0.39	0.03	0.64	0.46	0.03	0.85	0.26	0.03
a_2	0.69	0.47	0.03	0.63	0.5	0.03	0.85	0.26	0.03

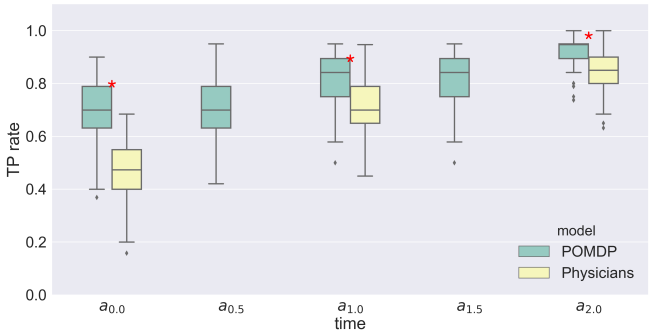
TABLE XVII: POMDP model performance using the DBN, an Exhaustive search model (all combinations of observations), and a logistic regression model, respectively. Testing data.

COMPARISON OF POMDP AND DBN:

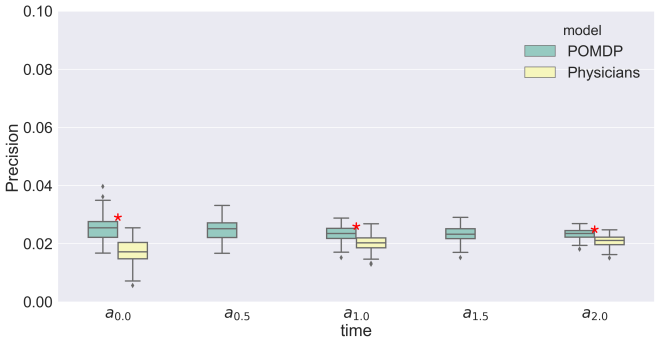
POMDP PERFORMANCE - COMPARISON OF OBSERVATION MODELS



(a) TN rate.



(b) Recall/TP rate.



(c) Precision.

Fig. 6: Box plots of the performance (TN, TP, precision) of the POMDP and physicians on cases from the NLST testing set, from the start of the trial through to last screening. Blue and yellow represent the POMDP and experts, respectively. Red stars depict instances where the performance measure between the physicians and model are significantly different.

BOX PLOTS OF ALL CASES: