Journal of Biomedical Informatics 68 (2017) 50–57

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Dynamic strategy for personalized medicine: An application to metastatic breast cancer

Xi Chen [a,*], Ross D. Shachter [a], Allison W. Kurian [b], Daniel L. Rubin [c]

[a] Department of Management Science & Engineering, Stanford University, Stanford, CA, USA
[b] Department of Medicine, Stanford University, Stanford, CA, USA
[c] Department of Radiology, Stanford University, Stanford, CA, USA

ARTICLE INFO

Article history:
Received 18 August 2016
Revised 30 January 2017
Accepted 18 February 2017
Available online 21 February 2017

Keywords:
Breast cancer
Dynamic treatment strategy
Personalized medicine
Markov Decision Process

ABSTRACT

We compare methods to develop an adaptive strategy for therapy choice in a class of breast cancer patients, as an example of approaches to personalize therapies for individual characteristics and each patient's response to therapy. Our model maintains a Markov belief about the effectiveness of the different therapies and updates it as therapies are administered and tumor images are observed, reflecting tumor response. We compare three different approximate methods to solve our analytical model against standard medical practice and show significant potential benefit of the computed dynamic strategies to limit tumor growth and to reduce the number of time periods patients are given chemotherapy, with its attendant side effects.

1. Introduction

Personalized medicine offers the potential of selecting the best therapies depending on patient characteristics, medical history, and observed response to treatment. We seek to develop a framework to model the effectiveness of different therapies and develop a strategy tailored to a patient or class of patients. In this paper we consider breast cancer patients who are hormone receptor-positive and we use approximate algorithms to construct therapy strategies that incorporate clinical observations about tumor response to therapy.

Breast cancer is one of the most common cancers with 231,840 estimated new cases and 40,290 estimated deaths among US women in 2015 [1]. Roughly 6% of all breast cancer cases reported from 2005 to 2011 are metastatic breast cancer cases. Breast cancer can be *hormone receptor-positive* for estrogen (ER+) and/or progesterone (PR+), where hormone therapy is most effective. This is the most prevalent of the three therapeutic categories for breast cancer, comprising 2/3 of all cases. Recent studies have also shown that targeted therapy in combination with hormone therapy, i.e. everolimus plus exemestane, is even more effective than many hormone therapies and chemotherapies [2,3]. Because hormone therapy, as well as targeted therapy, induce fewer side effects than chemotherapy, it is usually the first-line therapy for hormone receptor-positive breast cancer patients.

Oncologists face challenging questions when treating metastatic hormone-receptor positive patients, including

(1) ***Is the current therapy effective?***
We say that a therapy is currently *effective* for a patient if it is more likely that the tumor size will decrease than increase. According to a systematic review of metastatic breast cancer therapies [4], the rate of *objective tumor response* (defined as the tumor being less than half its initial size for at least 4 weeks), varies from 19% to 56% across therapies. There is considerable uncertainty, in the measurement of tumor size from radiological images, in the response of a tumor even to an effective therapy, and in the high probability that the measurement of tumor size will not change significantly in the two weeks between hospital visits. Together this makes it difficult to determine whether the current therapy is effective. This is exacerbated by the current practice of waiting three months between mammograms.
Unfortunately, tumors can develop resistance to therapies. The probability that a particular therapy is effective declines over time. This decline in effectiveness is observed whenever the patient receives that therapy or, to a lesser extent, another therapy from the same family, with similar functional mechanisms.
Therefore, we may possibly use the past treatment and observation history to update our belief about the current therapy effectiveness. However, this inference is too complicated to perform without a computer-based framework.

* Corresponding author.
E-mail address: besschenxi@gmail.com (X. Chen).

http://dx.doi.org/10.1016/j.jbi.2017.02.012
1532-0464/

(2) ***When to switch to another therapy?***
As the effectiveness of the current therapy is not directly observable, oncologists are unsure when to switch to other therapies. In the standard approach, each therapy is halted when the tumor *progresses* (defined as an increase of at least 25% in the estimated cross-sectional image area). However, evidence suggests that progression might not be the best signal to determine when to switch therapies. For example, early breast cancer patients who took Tamoxifen for 2–3 years and switched to Anastrozole have longer disease-free and local recurrence-free survival than those who continued with Tamoxifen [5].
Therefore, we seek to determine a better dynamic therapy strategy for hormone receptor-positive breast cancer patients so as to limit the tumor growth and reduce the chemotherapy side effects.

In the following sections: we build a belief Markov Decision Process (MDP) model for the therapy strategy; we validate model inputs and implement two Kalman Filter Q-Learning Algorithms, and compare numerical results for those algorithms, a generic solver SARSOP, and the standard medical practice; and we present our discussion and possible future work.

2. Methods

2.1. Decision framework for dynamic therapy strategy

We develop a dynamic decision framework to find an optimal strategy personalized to the patient. We choose to adopt Matlab for code development.

In our model, we consider alternative therapies for a patient with metastatic hormone receptor-positive breast cancer. There are two hormone therapies, Tamoxifen and Fulvestrant, and two chemotherapies, Docetaxel and Capecitabine, that we label as Therapy A, B, C and D, respectively. One of the four therapies can be chosen each period.

If we know which therapies are currently effective, we can plan a patient's treatment. A Markov Decision Process (MDP) models sequential decision problems with an underlying state. When the state is not directly observable the model becomes a Partially Observable Markov Decision Process (POMDP) [6]. A POMDP can then become a "belief MDP" with the introduction of a Markovian "belief state," the probability distribution for the underlying, but unobservable state. Our models use a Markov state to represent our current probabilities for effectiveness of the different therapies.

We estimated the prior belief of therapy effectiveness based on the literature, but in practice those could be adjusted for patient characteristics, eg, demographic, biological or microbiological. We assume that the effectiveness of different therapies are dependent if they are in the same family, i.e. hormone therapy or chemotherapy, but independent if they are in different families.

We use a simple Markov model to capture the decline of therapy effectiveness when applying the same therapy or a therapy in the same family. When therapy x is applied for a period there are two possible effects: the within-therapy effectiveness decline rate α_x denotes the probability that therapy x becomes ineffective, and the between-therapy effectiveness decline rate $\beta_{y|x}$ denotes the probability that therapy y in the same family becomes ineffective. To illustrate, if Tamoxifen (therapy A) is applied for one period, the effectiveness of hormone therapies A and B can change, as represented as in Fig. 1, but we assume that the effectiveness of chemotherapies C and D do not.

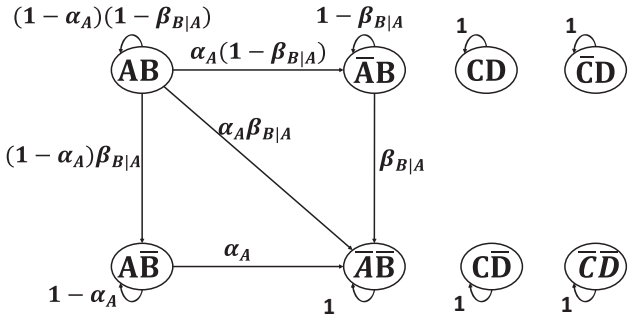

Fig. 1. Markov model of the changes to the effectiveness of therapies given therapy A is administered this period.

We use an MDP to model dynamic therapy choices. The length of each period corresponds to the time between hospital visits, when therapy decisions are usually made. For metastatic breast cancer patients, we assume periods last two weeks, and we use a time horizon of 100 periods or about 4 years. While the effectiveness of therapies are useful distinctions, they are not directly observable, so instead our state includes our beliefs about them, represented as a joint probability distribution. The state also includes the estimated tumor size, and we use it to update our beliefs about effectiveness.

For each period t, we define absolute tumor size m_t in square centimeters as the estimate obtained by summing over the product of the two dimensions of observable tumors given in the radiological report that period. For example, the initial tumor size is m_0. We believe that the tumor size in period $t+1$, m_{t+1}, will depend on the therapy chosen, its effectiveness, and m_t. A therapy is chosen in period t after observing m_t and updating our beliefs about the current effectiveness of the therapies. The stage reward in period t combines the tumor response and therapy side effects, and is a function of the therapy chosen and m_t, discounted at the rate of 0.95 per year. The first three time periods of the model are represented by the influence diagram shown in Fig. 2. The dashed boxes group the therapy effectiveness uncertainties in each period, representing that arcs directed into and out from the dashed box apply to every uncertainty in the dashed box.

We could have solved this problem as a discrete state POMDP, but since we would need to remember all past observations to update our belief, it becomes intractable as the number of therapies and observations increase. Instead, we solve a "belief state" MDP, characterized by its states, actions, and stage rewards, as described below.

Belief MDP Model

Action: In each period, we choose a therapy $a \in \{A, B, C, D\}$.

Reward: Stage reward is separated into two parts, to be summed and discounted at the rate of 0.95 per year.

The first part is the tumor response, which we define as the negative log of the tumor size, $-\log_2(m_t)$ for $t = 1, \ldots, 100$ and $-\log_2(m_{100})$ thereafter, corresponding to the MDP shown in Fig. 2. Since we have no control over the initial size m_0, it suffices to minimize the number of tumor size doublings,

$$R_{1,t} = -\log_2\left(\frac{m_t}{m_{t-1}}\right) \text{ for all } t = 1, \ldots, 100, \tag{1}$$

corresponding to the MDP shown in Fig. 3. We assume that the observed tumor response will grow by one of three factors in period t: either *increase* to $2^{1/6} \approx 112\%$, *decrease* to $2^{-1/6} \approx 89\%$, or *stay the same*, relative to the start of period t. We measure tumor size increases by factors of $2^{1/6} \approx 112\%$ to capture tumor progression, defined as a 25% increase in cross-sectional image area. Specifically,

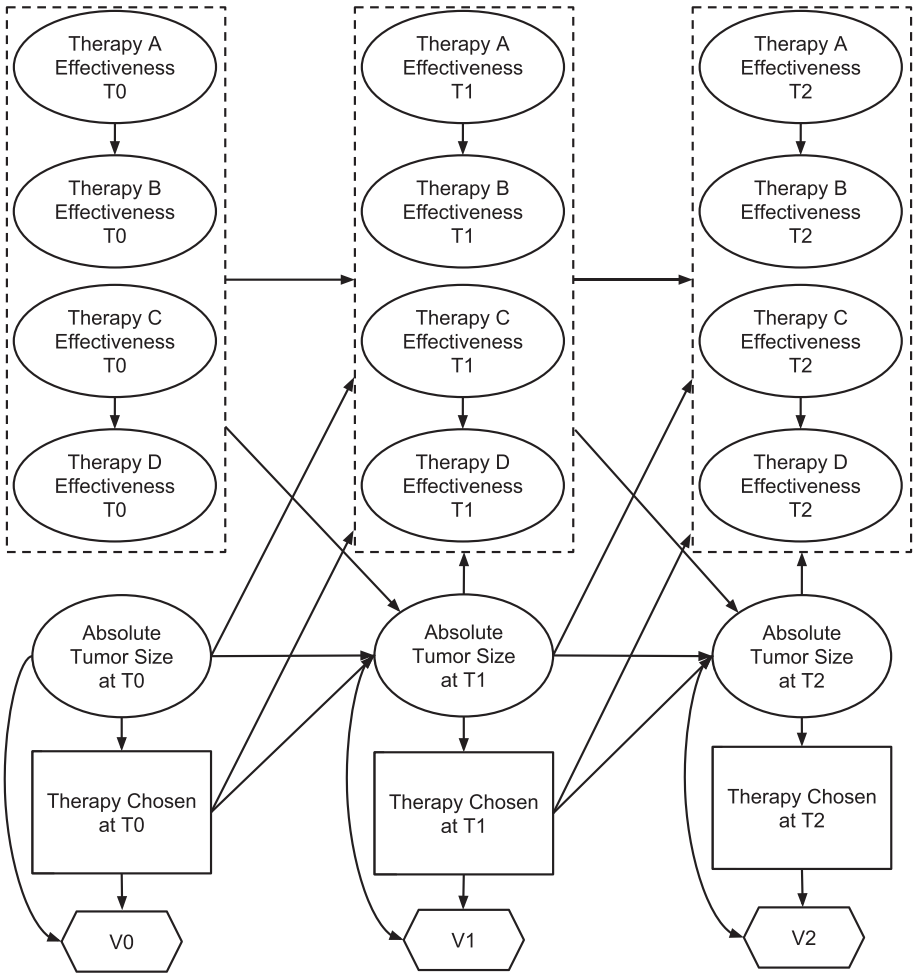

Fig. 2. MDP with state including current and past absolute tumor size and our beliefs about the effectiveness of therapies.

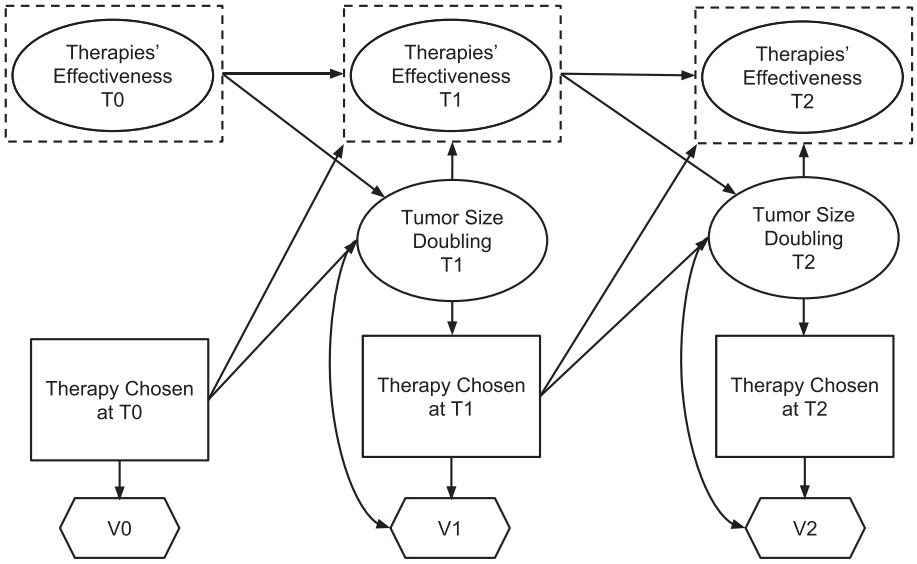

Fig. 3. MDP with state including current tumor response and our beliefs about the effectiveness of therapies.

two consecutive tumor size increase by factor of $2^{1/6}$ is equivalent to $2^{1/3} \approx 125\%$.

The second part of stage reward represents the disutility due to chemotherapy side effects. We define $\phi = -0.1$ as the value that makes the decision maker indifferent between (1) enduring the side effects caused by chemotherapy for $1/|\phi| = 10$ periods, i.e. 20 weeks, or (2) having the tumor become twice as big as with that chemotherapy. We assume that both Docetaxel and Capecitabine,

i.e. therapy C and D, have the same disutility due to chemotherapy side effects in our model. Therefore, the chemotherapy disutility is

$$R_{2,t}(a) = \begin{cases} \phi = -0.1 & \text{if } a \in \{C, D\} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The total stage reward is $R_t = 0.95^{t/26}(R_{1,t} + R_{2,t})$.

State: The current state s is the current tumor response and joint probability of effectiveness for all therapies, accounting for the dependence between therapies in the same family,

$$s = (R_{1,t}, p_A, p_{B|A}, p_{B|\overline{A}}, p_C, p_{D|C}, p_{D|\overline{C}}).$$

Given the state s and action a, we update to obtain our next state s'.

2.2. Model Input Validation

In the belief MDP shown in Fig. 3, four sets of inputs are required, namely the tumor response observation probability matrix, the initial probability of effectiveness, within-therapy decline rate α and between-therapy decline rate β. We assume that the tumor response observation probability matrix is the same for all therapies, whereas the other three inputs are assessed for each therapy.

The ***tumor response observation probability matrix*** describes the probability of tumor grows, shrinks or stays the same given an effective or ineffective therapy is administrated for the period. In the influence diagram shown in Fig. 3, this probability matrix is embedded in the *Tumor Size Doubling* uncertainty. The tumor response observation probability matrix we used is shown in Table 2. As a first step to assess model inputs, we assume this probability matrix is based on the three degrees of tumor response, i.e. either increasing to $2^{1/6} \approx 112\%$, decreasing to $2^{-1/6} \approx 89\%$, or staying the same.

We assess the ***initial probability of effectiveness*** p_i ***and within-therapy decline rate*** α for each therapy individually based on the assumed tumor response observation probability matrix. We refer to clinical measures reported in randomized clinical trial results where each therapy is applied individually as a first-line therapy. The clinical measures include *Progression Free Survival* and the *clinical benefit rate*. PFS is defined as the time from randomization to the trial until tumor progression or death, usually measured in months. Clinical benefit rate is the percentage of patients with objective tumor response or *stable disease* (a decrease in the cross sectional image area maintained for at least 6 months).

For each therapy, we extract three numbers from clinical trials, namely the median PFS, the 80th percentile of PFS, and the clinical benefit rate. Then we perform grid search over pairs of initial probability of effectiveness and α to identify the pair that minimizes the square error in comparison to the three randomized clinical trial statistics.

The initial probability of effectiveness p_i we obtain from parameter fitting is the marginal probability for individual therapy. The dependence of effectiveness between therapies is difficult to estimate from clinical trial results, and we relied on subjective judgments. In this research, we assess the conditional probabilities, and ensure they are consistent with the marginal probabilities from parameter fitting.

The ***between-therapy decline rate*** β is even harder to estimate from randomized clinical trial data. To approximate β , suppose that the clinical trial statistics for the same therapy i as a first-line versus second-line therapy (given therapy j is used as a first-line therapy) are both available. We can perform similar grid search to find the optimal pair of initial probability of effectiveness p_i and α for both cases. Let p_{i1} denote the initial probability of effectiveness when therapy i is used as a first-line therapy; and p_{i2} denote the initial probability of effectiveness when therapy i is used as a second-line therapy after t periods of therapy j. Then we can estimate β_{ij} such that $p_{i2} = p_{i1}(1 - \beta_{ij})^t$.

For example, we estimate between-therapy effectiveness decline rate for hormone therapies using clinical trial data showing the effectiveness of therapy i = Fulvestrant following therapy j = Tamoxifen that can be compared to starting out with therapy i.

2.3. Kalman Filter Q-Learning Methods

Our solution is a (near) optimal policy mapping from the continuous belief state to action, usually with an approximate algorithm. In particular, the exact and approximate Kalman Filter Q-learning (KFQL and AKFQL) Methods [7] search for an effective policy for multi-dimensional continuous state space MDP's via Q-Learning [8]. We applied them to our belief MDP. Definition of variables in KFQL and AKFQL methods are listed in Table 1.

Given a set of basis functions $\theta(s, a)$ for state s and action a, we learn a vector of weights r for these basis functions to minimize the mean Bellman residual. Prior belief is that r follows a multivariate normal distribution with mean μ and covariance matrix Σ. Q-value, calculated from the basis functions and the weights r, approximates the optimal net present value of the future rewards given initial state action pair (s, a).

$$Q(s, a) = r^T \theta(s, a) \quad (3)$$

$$\approx R(s, a, s') + \gamma \max_{a'} Q(s', a') \quad (4)$$

When updating the weights r using a Kalman filter, each sample has variance $\epsilon(s, a)$, assumed to be conditionally independent of the prediction given r.

AKFQL is a simplified version of KFQL that ignores the dependence among basis functions, treating Σ as a diagonal matrix. AKFQL reduces the computation complexity from quadratic to linear in the number of basis functions. It also appears to be more efficient and robust generating policies for our problem.

We chose two basis functions for each therapy, so 8 basis functions are used for our 4 therapies. For the therapy chosen, they are the probability it is effective p_i and a constant 1; they are both zero for the therapies not chosen. Symbolically, the two basis functions corresponding to each therapy i are given by

$$\theta_{i,1} = \begin{cases} p_i & \text{if therapy } i \text{ is chosen} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$\theta_{i,2} = \begin{cases} 1 & \text{if therapy } i \text{ is chosen} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Both KFQL and AKFQL require an estimate of estimation error or sensor noise, ϵ, used to control the learning rate. We found $\epsilon_0 = 20$ for KFQL and $\epsilon_0 = 1$ for AKFQL worked well in our model. We found the best method to compute ϵ for KFQL was the "average method," averaging over all alternatives, and the best method for AKFQL was the "policy method," based on the optimal choice. In

Table 1
Definition of variables in (approximate) Kalman Filter Q-Learning Method.

Variable	Definition
s	Current MDP state
s'	Successor MDP state
a	Action, therapy choice from {A, B, C, D}
$R(s, a, s')$	MDP reward
$\theta(s, a)$	Basis function values for pair (s, a)
r	Weights on the basis functions that minimize the mean Bellman residual
μ, Σ	Mean vector and covariance matrix for $r \sim N(\mu, \Sigma)$
$Q(s, a)$	Q-value for (s, a)
$\epsilon(s, a)$	Estimation error, learning rate

both algorithms, we adopt the backtracking technique to update the policy [9].

For comparison, we also applied a POMDP solver using a point-based algorithm. Point-Based Value Iteration (PBVI) [10] was the first point-based algorithm that demonstrated good performance on large state space POMDPs. Based on the understanding that most POMDP problems are unlikely to reach most of the belief states even with arbitrary action and observation sequences, PBVI selects a small set of representative belief states and performs value updates to these states iteratively. Many more point-based algorithms have been developed with improved performance [11,12]. We chose to apply Successive Approximations of the Reachable Space under Optimal Policies (SARSOP) [13] to our problem. SARSOP explicitly attempts to sample the optimally reachable belief states through learning-enhanced exploration and a bounding technique. It maintains a belief tree, with the initial belief state as its root, and prunes out subtrees that will never be visited under the optimal policy.

3. Results

3.1. Model Inputs

Probability inputs for our model are validated against the medical literature. A paper which conducted random trials comparing Tamoxifen and Fulvestrant is used to validate inputs for Therapy A and B [14]. Similarly, papers with clinical trials for Docetaxel and Capecitabine are used for Therapy C and D [15,16]. We also compared the hormone therapy used with and without hormone therapy history to infer the between-therapy effectiveness decline rate β [17].

Table 2 shows how the tumor response we observe depends probabilistically on therapy effectiveness. Table 3 lists the parameters used in the model, and compares the model outputs to the statistics from randomized clinical trials where exactly one therapy was used. Although the side effects from chemotherapy may indirectly affect PFS and clinical benefit, they are not explicitly considered.

Table 2
Tumor response observation probability.

	Increase	Stay the same	Decrease
If therapy is effective	0.1	0.5	0.4
If therapy is ineffective	0.8	0.2	0

Table 3
Model input validation.

		Tamoxifen	Fulvestrant
P(Effective)		90.5%	88%
α		4.5%	5.5%
β		1.5%	1.5%
Median	**Model**	8.86	7
PFS (mo.)	**Trial**	8.3	6.8
80th %tile	**Model**	21.93	17.73
PFS (mo.)	**Trial**	22	18
Clinical	**Model**	62.9%	55.7%
Benefit Rate	**Trial**	62%	54.3%
		Docetaxel	**Capecitabine**
P(Effective)		97.5%	82.5%
α		6.5%	6.5%
β		2%	2%
Median	**Model**	7.47	6.07
PFS (mo.)	**Trial**	7.5	6
80th %tile	**Model**	16.3	14.47
PFS (mo.)	**Trial**	15	15
Clinical	**Model**	57.4%	49.0%
Benefit Rate	**Trial**	58%	50%

3.2. Convergence to optimal policy

As mentioned before, the Kalman Filter Q-Learning algorithms are iterative methods to find an approximate optimal policy. We used a time horizon of 100 periods to provide meaningful strategies for at least four years of therapy. We found that both algorithms converged with 100 independent sample paths. The complexities of KFQL and AKFQL are quadratic and linear, respectively, in the number of basis functions, two for each therapy.

In order to investigate convergence, we evaluated the policies generated using samples with different time horizons. Specifically, 100 random samples with time horizon 100 periods are first generated independently. Then, the algorithms are applied separately to the samples of the first period, the first 2 periods, and eventually all 100 periods. The resulting policies are recorded and then evaluated by simulation of 15,000 independent samples and the average total reward over 100 periods is calculated. The resulting average rewards are then plotted on the same diagram shown in Fig. 4. For 100 samples and 8 basis functions, both methods seemed to converge within the time horizon, and both methods converge to policies with similar average reward. More interestingly, AKFQL, which ignores the covariance between coefficients, converges faster than KFQL.

3.3. Comparison of policies

We compared the policy generated by the Kalman Filter Q-Learning methods to another approximate algorithm and standard medical practice.

The approximate POMDP solver SARSOP [13] found a policy with average 100 period reward of -5.96, with 95% confidence interval $(-6.05, -5.87)$.

The standard therapy strategy uses hormone therapies before chemotherapies, and switches therapies when progression is observed, i.e., tumor size increases at least 25% over its value at the start of the therapy. Specifically, we assumed that the sequence of therapies is $A, B, C, D, \varnothing$. If all the therapies have failed, therapy $\varnothing$ indicates no therapy and disease progresses as if there were an ineffective therapy without side effects. The standard therapy strategy results in an average total value of -6.47, with 95% credible interval $(-6.58, -6.36)$.

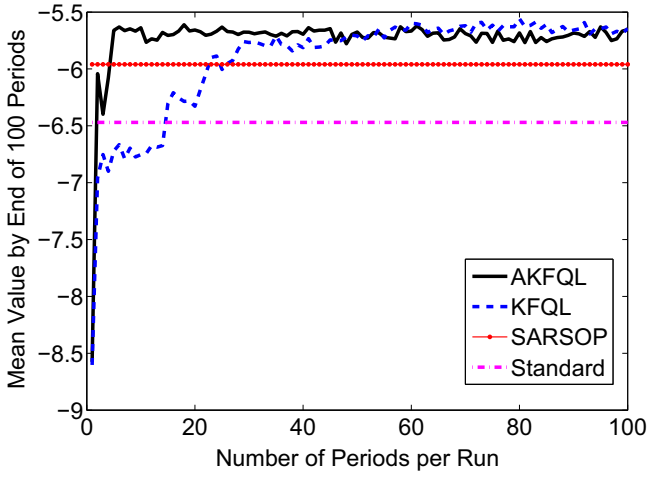

Fig. 4. KFQL and AKFQL converge to policies with similar average rewards, but AKFQL converges faster. They both generate policies that achieve higher average reward than SARSOP and the standard therapy strategy.

According to the model, all three methods for solving this POMDP appear to yield significantly better results than standard medical practice. AKFQL and KFQL achieve the best results and AKFQL does so with the least computational complexity.

As the reward trades off the tumor response (or shrinkage) against the disutility from chemotherapy side effects, we considered these two elements separately for the four policies. Fig. 5 shows the mean of the cumulative number of tumor doublings under all four policies over time. The y-axis is the number of tumor doublings, the logarithm of relative tumor size, $\log_2(m_t/m_0)$. Fig. 6 shows the mean cumulative number of periods that chemotherapies, i.e. therapy C or D, are used under each policy.

The policies from KFQL and AKFQL perform similarly for most of the measurements. Their policies are the most effective at limiting tumor growth. Compared to the standard therapy strategy, there is half as much median tumor growth under their policies at period 40. Moreover, their policies apply less chemotherapy than the standard therapy strategy, on average, 22.03 and 23.97 v.s. 28.57 out of 100 periods. Thus, it appears that they dominate the standard therapy strategy under our model assumptions.

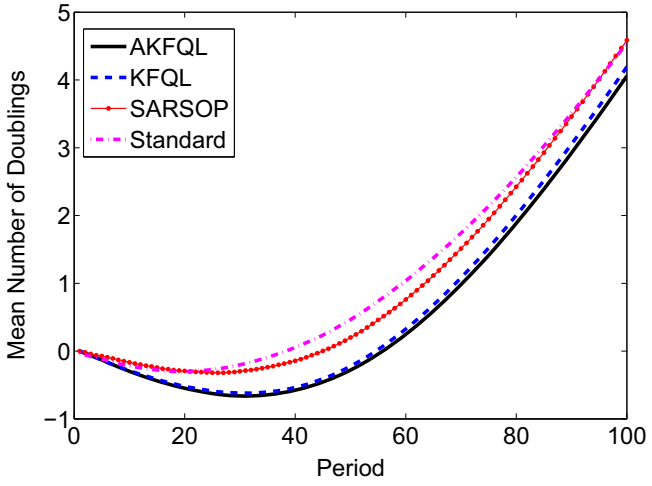

Fig. 5. The cumulative number of tumor doublings shows that the policies from KFQL and AKFQL are the most effective at limiting tumor growth.

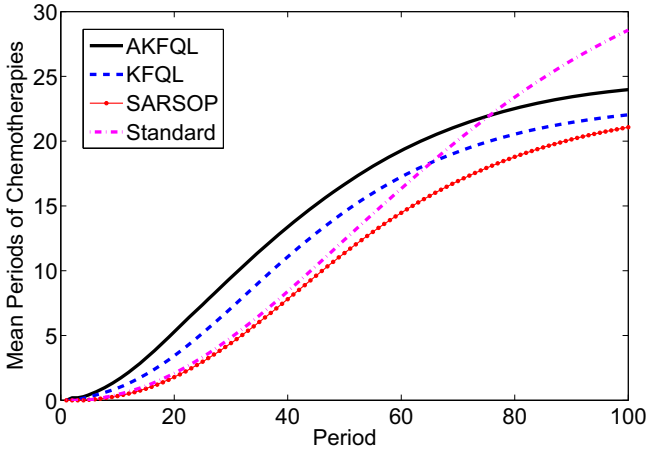

Fig. 6. The cumulative number of periods of chemotherapies shows that all three computed policies apply chemotherapy in fewer periods than the standard therapy strategy.

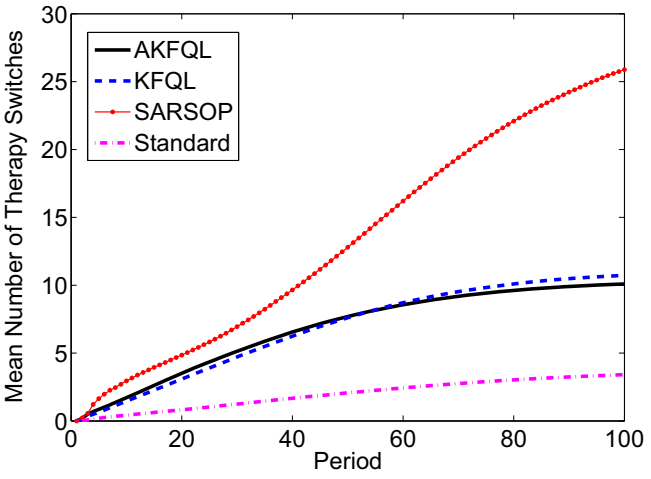

Fig. 7. The cumulative number of therapy switches varies drastically across the four policies. The more frequent switches in the computed policies may lead to prolonged estimated PFS.

SARSOP's policy, in a similar way, limits tumor growth better than the standard therapy strategy while applying chemotherapy in the fewest periods among all therapy strategies, 21.07 out of 100 periods on average.

A significant difference between the standard therapy strategy and the three computed policies is the number of therapy switches, as shown in Fig. 7 and reflected in sample therapy trajectories shown in Fig. 8.[1] The standard therapy strategy iterates once through all four therapies, from hormone therapies to chemotherapies, using progression as the signal to switch, and we estimate an average of 3.41 switches in 100 periods, with progression under all four therapies in most samples, leading to cessation of therapy. On the other hand side, the policies generated by KFQL, AKFQL, and SARSOP have on average 10.73, 10.06, and 25.89 therapy switches in 100 periods. The significantly higher number of therapy switches in SARSOP's policy suggests an overly complicated, or overfitted, policy.

We speculate that this more frequent switch of therapy contributes to the better performance of these three policies as it enables the therapies to be applied when our model considers them more effective. Due to the loss of effectiveness over time, especially from therapies in the same family, therapies are less likely to be effective when applied later. For example, if therapy B is only applied when therapy A results in tumor progression, therapy B may itself no longer be effective even if it would have been earlier.

We estimated the clinical measures for all four policies using simulation, as shown in Table 4. These measures do not reflect the patient's disutility from chemotherapy. The policies from KFQL and AKFQL result in much better performance in both PFS and clinical benefit rate than SARSOP's policy and the standard therapy strategy, which corresponds to the effective control of tumor size. SARSOP's policy results in a clinical benefit rate similar to the standard therapy strategy, but with much longer PFS.

We also plot the progression-free survival curves for these four policies in Fig. 9. More than 60% of all patients are estimated to have progression-free survival longer than 20 months under AKFQL and KFQL policies, compared to roughly 20% under the standard policy. However, AKFQL and KFQL policies apply chemotherapies for an average of 2.5 months more than the Standard policy during the first 20 months, as shown in Fig. 6.

[1] Fig. 8 demonstrates that given the same initial beliefs about therapy effectivenesses, therapies should be chosen differently when different tumor responses are observed. Each line represents one possible treatment trajectory from period 1–100.

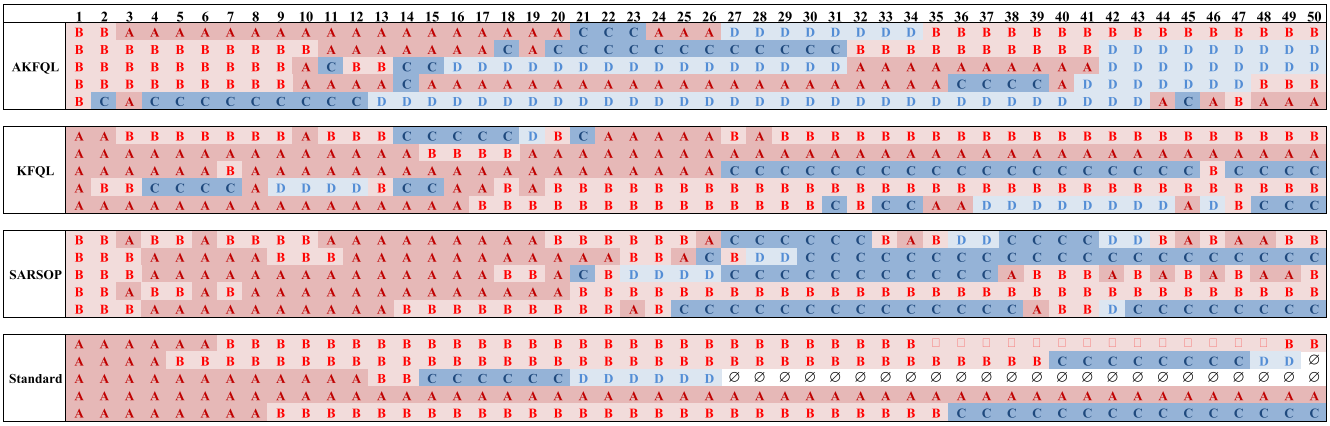

Fig. 8. Sample therapy trajectories in the four policies show how the patient may need to follow different therapy trajectories when the observed tumor sizes differ.

Table 4
Efficacy of the different policies.

Method	Median PFS (mo.)	Clinical benefit rate (%)	Periods of chemotherapy
KFQL	27.07	88.47	22.02
AKFQL	28.47	90.56	23.97
SARSOP	17.73	79.7	21.07
Standard	8.87	79.2	28.57

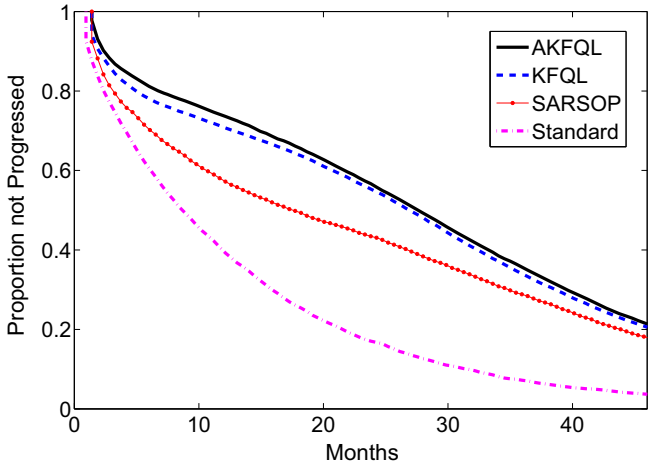

Fig. 9. The cumulative number of therapy switches varies drastically across the four policies. The more frequent switches in the computed policies may lead to prolonged estimated PFS.

Futhermore, all three computed policies apply chemotherapy in fewer periods than the standard therapy strategy. According to the these policies, therapy switches are executed before progression is observed. This may have contributed to the prolonged PFS.

3.4. Sensitivity analysis of chemotherapy toxicity factor ϕ

We performed sensitivity analysis on the chemotherapy toxicity factor ϕ, which we assumed to be -0.1 in our analysis, because it incorporates assumptions about patient preferences and we expect it to be different for different patients. Holding all other model parameters fixed, we varied ϕ from -0.3 to 0, in 0.01 increments. When $\phi = 0$, the patient has zero disutility for chemotherapy side effects; thus the only reward measure is tumor response. As we increase the disutility, i.e. ϕ becomes more negative, the patient prefers less chemotherapy (see Fig. 10).

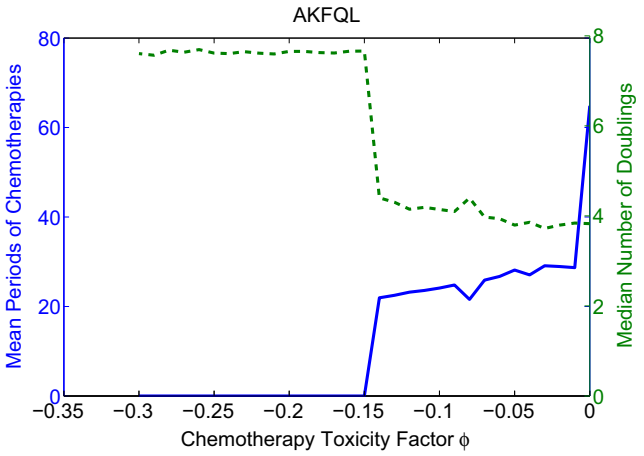

Fig. 10. The mean number of periods of chemotherapies and the median number of tumor doublings vary gradually with chemotherapy toxicity factor ϕ between -0.14 and -0.01; for $\phi < -0.15$, chemotherapy is not applied; as ϕ approaches 0, there is an increase in chemotherapy use, but no reduction in tumor doubling.

When the chemotherapy toxicity factor ϕ is less than or equal to -0.15, the approximate optimal policies do not apply chemotherapies, i.e. therapy C and D. As ϕ increases from -0.15 to -0.01, the median number of tumor doublings decreases significantly from 7.7 to 4.4 while the mean number of periods of chemotherapy in the optimal policies increases gradually from 21.9 to 28.7 periods. Finally, as ϕ approaches 0, the mean number of periods of chemotherapies increases dramatically to 64.9 periods, without any significant change in the median number of tumor doublings.

4. Discussion

We developed therapy strategies for a hormone receptor-positive breast cancer patient that outperform standard medical practice, according to our model, but they would have to be confirmed by prospective patient studies. All of our computed strategies, based on belief Markov Decision Process (MDPs), appear to prolong progression free survival and increase clinical benefit rate with less chemotherapy than the standard therapy strategy. They try to switch therapies before the tumor progresses and the therapy, as well as alternative therapies in the same family, lose their effectiveness.

Although the POMDP Solver SARSOP outperformed the standard therapy, its strategy was more complicated and less effective than those generated by Kalman Filter Q-Learning (KFQL) and

Approximate KFQL (AKFQL). AKFQL is a simplified version of KFQL which is considerably faster and converged more rapidly to its final policy. Both AKFQL and KFQL appear to be promising techniques for solving belief MDPs, where the system maintains a multi-dimensional continuous joint probability as a Markov state.

Our framework can be personalized to patients with different chemotherapy toxicity disutility, as shown in the sensitivity analysis in the preceding section. We also tested our framework using more than four therapies, and showed that it scales well with the number of alternative therapies and the resulting increased state space, especially when applying the AKFQL method.

As personalized medicine develops and the number of highly-targeted alternative therapy increases, there is a need for sophisticated strategies to manage each patient's therapy, taking into account the patient's characteristics, medical history, observed tumor response, and side effects. With increasing patient outcomes data available, the belief MDP model and AKFQL solution method might be able to develop a customized therapy strategy that can be adapted in real time to advise patients and physicians. This same approach could also be applied in other medical and non-medical domains.

There are promising directions to further develop this research effort, which would either advance our effort to develop dynamic strategies for metastatic breast cancer patients, or extend this integrated framework to other domains.

First of all, we have assumed that the chemotherapy toxicity factor is constant over time in our model. In other words, the cumulative chemotherapy toxicity is linear in time. However, research has found that even though the cumulative dose of chemotherapy is the most robust risk factor of toxicity, the relationship is not necessarily linear. As an example, for chemotherapy doxorubicin, the estimated percentage of patients with doxorubicin-related heart failure was found to be 5% at a cumulative dose of 400 mg/m^2, 26% at 550 mg/m^2, and 48% at 700 mg/m^2 [18].

To better capture the toxicity factor, we can include the exponentially smoothed cumulative dose of applying a therapy as a Markov state variable, or even more specifically the cumulative toxicity of a therapy. Some questions may need to be addressed in the process. For example, how to capture the decrease in cumulative toxicity when a therapy is paused for some periods; how to make personalized assessments for cumulative toxicity as a function of dosages over time, etc. With more clinical knowledge of therapy toxicity, we may improve our model of toxicity.

Secondly, we can use our model to include combination therapy. Each combination therapy regimen will be considered as one therapy in our model, with an assigned probability of effectiveness, and with-in therapy decline rate α. When assigning between-therapy decline rate β's, we will take into consideration any beliefs about the resistance tumor may develop to other therapies, including each single agent included in the combination therapy.

Thirdly, we have used single studies for the four therapy alternatives selected. There are multiple clinical trials available for each therapy, with marginally or significantly different reported effectiveness [19,20]. Thus, more sophisticated methods should be used to aggregate data from different clinical trials to set input parameters of the model. Moreover, actual results in the real world setting may differ from the clinical trial results. One possible mitigation method would be to incorporate expert judgment from oncologists when setting model inputs.

Last but not the least, there is potential that our integrated framework could be adopted for other problems in medical or non-medical domains. Examples include treatment strategies of other types of cancer, e.g. lung cancer, lymphoma, etc, treatment strategy of infectious disease, and economic problems like dynamic pricing [21] and optimized marketing [22].

References

[1] Natinoal Cancer Institute, SEER Cancer Statistics Factsheets: Female Breast Cancer, 2014.

[2] Thomas Bachelot, Rachael McCool, Steven Duffy, Julie Glanville, Danielle Varley, Kelly Fleetwood, Jie Zhang, Guy Jerusalem, Comparative efficacy of everolimus plus exemestane versus fulvestrant for hormone-receptor-positive advanced breast cancer following progression/recurrence after endocrine therapy: a network meta-analysis, Breast Cancer Res. Treat. 143 (1) (2014) 125–133.

[3] Daniele Generali, Sergio Venturini, Carla Rognoni, Oriana Ciani, Lajos Pusztai, Sherene Loi, Guy Jerusalem, Alberto Bottini, Rosanna Tarricone, A network meta-analysis of everolimus plus exemestane versus chemotherapy in the first- and second-line treatment of estrogen receptor-positive metastatic breast cancer, Breast Cancer Res. Treat. 152 (1) (2015) 95–117.

[4] R. Fossati, C. Confalonieri, V. Torri, E. Ghislandi, A. Penna, V. Pistotti, A. Tinazzi, A. Liberati, Cytotoxic and hormonal treatment for metastatic breast cancer: a systematic review of published randomized trials involving 31,510 women, J. Clin. Oncol. 16 (10) (1998) 3439–3460.

[5] F. Boccardo, Switching to anastrozole versus continued tamoxifen treatment of early breast cancer: preliminary results of the Italian Tamoxifen Anastrozole Trial, J. Clin. Oncol. 23 (22) (2005) 5138–5147.

[6] R.D. Smallwood, E.J. Sondik, The Optimal Control of Partially Observable Markov Decision Processes over a Finite Horizon, 1973.

[7] Charles Tripp, Ross Shachter, Approximate Kalman Filter Q-Learning for Continuous State-Space MDPs, in: Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13), 2013, pp. 644–653.

[8] C J C H. Watkins, Learning from Delayed Rewards (Phd Thesis), University of Vambridge, England, 1989.

[9] Charles Tripp, Ross Shachter, Backtracking for more efficient large scale dynamic programming, in: 2012 11th International Conference on Machine Learning and Applications, 2012, pp. 338–343.

[10] Joelle Pineau, Geoff Gordon, Sebastian Thrun, Point-based value iteration: an anytime algorithm for POMDPs, in: IJCAI International Joint Conference on Artificial Intelligence, 2003, pp. 1025–1030.

[11] Guy Shani, Joelle Pineau, Robert Kaplow, A survey of point-based POMDP solvers, Auton. Agent. Multi-Agent Syst. 27 (1) (2013) 1–51.

[12] Trey Smith, Reid Simmons, Heuristic Search Value Iteration for POMDPs, Uai, 2004, pp. 520–527.

[13] Hanna Kurniawati, David Hsu, Wee Sun Lee, SARSOP: efficient point-based POMDP planning by approximating optimally reachable belief spaces, in: Proceedings of Robotics: Science and Systems IV, 2008, page w/o page numbers.

[14] A. Howell, J.F.R. Robertson, P. Abram, M.R. Lichinitser, R. Elledge, E. Bajetta, T. Watanabe, C. Morris, A. Webster, I. Dimery, C.K. Osborne, Comparison of fulvestrant versus tamoxifen for the treatment of advanced breast cancer in postmenopausal women previously untreated with endocrine therapy: a multinational, double-blind, randomized trial, J. Clin. Oncol. 22 (9) (2004) 1605–1613.

[15] William J. Gradishar, Dimitry Krasnojon, Sergey Cheporov, Anatoly N. Makhson, Georgiy M. Manikhas, Alicia Clawson, Paul Bhar, Significantly longer progression-free survival with nab-paclitaxel compared with docetaxel as first-line therapy for metastatic breast cancer, J. Clin. Oncol. 27 (22) (2009) 3611–3619.

[16] M.R. Stockler, V.J. Harvey, P.A. Francis, M.J. Byrne, S.P. Ackland, B. Fitzharris, G. Van Hazel, N.R.C. Wilcken, P.S. Grimison, A.K. Nowak, M.C. Gainford, A. Fong, L. Paksec, T. Sourjina, D. Zannino, V. Gebski, R.J. Simes, J.F. Forbes, A.S. Coates, Capecitabine versus classical cyclophosphamide, methotrexate, and fluorouracil as first-line chemotherapy for advanced breast cancer, J. Clin. Oncol. 29 (34) (2011) 4498–4504.

[17] C.K. Osborne, Double-blind, randomized trial comparing the efficacy and tolerability of fulvestrant versus anastrozole in postmenopausal women with advanced breast cancer progressing on prior endocrine therapy: results of a North American trial, J. Clin. Oncol. 20 (16) (2002) 3386–3395.

[18] Antonis Valachis, Cecilia Nilsson, Dove Press, Cardiac risk in the treatment of breast cancer: assessment and management, Breast Cancer 7 (2015) 21–35 (Dove Medical Press).

[19] E. Campora, G. Colloca, R. Ratti, G. Addamo, Z. Coccorullo, A. Venturino, D. Guarneri, Docetaxel for metastatic breast cancer: two consecutive phase II trials, Anticancer Res. 28 (6 B) (2008) 3993–3995.

[20] H.J. Stemmler, K. Gutschow, H. Sommer, M. Malekmohammadi, C.H. Kentenich, R. Forstpointner, S. Geuenich, J. Bischoff, W. Hiddemann, V. Heinemann, Weekly docetaxel (Taxotere) in patients with metastatic breast cancer, Annals Oncol.: Official J. Eur. Soc. Medical Oncol./ ESMO 12 (10) (2001) 1393–1398.

[21] Paat Rusmevichientong, Joyce A. Salisbury, Lynn T. Truss, Benjamin van Roy, Peter W. Glynn, Opportunities and challenges in using online preference data for vehicle pricing: a case study at General Motors, J. Revenue Pricing Manage. 5 (1) (2006) 45–61.

[22] Naoki Abe, Naval Verma, Chid Apte, Robert Schroko, Cross Channel Optimized Marketing by Reinforcement Learning, in: The 2004 ACM SIGKDD International Conference, vol. 3, 2004, p. 767.