Crowdsourcing Reinforcement Learning to Optimize Knee Replacement Pathway

Hao Lu [1] **Mengdi Wang** [1]

Abstract

Joint replacement is the most common inpatient surgical treatment in the US. We investigate the clinical pathway optimization for knee replacement, which is a sequential decision process from onset to recovery. Based on episodic claims from previous cases, we view the pathway optimization as an intelligence crowdsourcing problem and learn the optimal decision policy from data by imitating the best expert at every intermediate state. We develop a reinforcement learning-based pipeline that uses value iteration, state compression, kernel representation and cross validation to predict the best treatment policy. It also provides forecast of the clinical pathway under the optimized policy. Empirical validation shows that the optimized policy achieves 7% reduction of the overall cost and 33% reduction of the excessive cost premium.

1. Introduction

Driven by the explosion of computing power and big data, the past decade has witnessed tremendous success of artificial intelligence (AI) in many applications, like image recognition (Russakovsky et al., 2015), natural language processing (Manning et al., 2014), game bot (Mnih et al., 2015), and speech recognition (Hinton et al., 2012). Meanwhile, the healthcare industry, with massive volumes of clinical data together with sophisticated devices and recording systems, stands to benefit from this new era of artificial intelligence. Machine learning technologies have proved effective for analyzing certain medical data and providing end-to-end predictions; see for examples (Esteva et al., 2017; Shickel et al., 2018; Raghu et al., 2017b).

However, artificial intelligence is way underdeveloped for clinical decision making in complex situations. People face tremendous challenges. Healthcare data are featured for high-dimensionality, heterogeneity, temporal dependency, and irregularity (Luo et al., 2016; Miotto et al., 2017; Doorhof, 2018). Machine learning methods can produce good predictions at certain scenarios, but their model often lacks interpretability, generalizability and global optimization. The healthcare system often involves ineffective services, excessive cost, and uncoordinated cares (Olsen et al., 2010). Conventional machine learning methods cannot prescribe every clinical decision in a systematic way. There lacks a principled AI-driven approach to optimize the full clinical decision process for complicated medical treatments.

In this paper, we focus on the clinical decision process of knee replacement. During an episode of knee replacement treatment, a patient has to go through a sequence of diagnosis, prescriptions and treatments, which we refer to as a clinical pathway. We develop a reinforcement learning approach to optimize the knee replacement pathway based on claims records. Empirical results demonstrate that our solution achieves significant improvement over the baseline.

1.1. Knee Replacements

Knee replacement is a surgical procedure using metal and plastic parts to resurface damaged knees. An episode of knee replacement treatment takes approximately three months. During this process, a patient go through a pathway from onset to surgery, rehabilitation and recovery. According to the Centers for Medicare and Medicaid Services (CMS) (CMS, 2016), hip and knee replacements are the most common inpatient surgery in the United States. Around 700,000 knee replacement procedures take place every year (Martin, 2017). The average cost per episode ranges from $16,500 to $33,000 across different geographical areas (CMS, 2016). Patients who receive the knee replacement treatments often have to go through a long recovery time. Knee replacement treatments are costly for the patients and the Medicare, and the quality of treatments vary largely among healthcare providers.

Beginning on April 1, 2016, the Centers for Medicare and Medicaid Services (CMS) started a new payment model,

[1]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA. Correspondence to: Mengdi Wang <mengdiw@princeton.edu>.

Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36th International Conference on Machine Learning, Long Beach, California, USA, 2019. Copyright 2019 by the author(s).

called the Comprehensive Care for Joint Replacement (CJR) model (CMS, 2016), to reduce excessive costs and support better care for patients who are undergoing elective hip and knee replacement surgeries. The CJR model will run for five years in 67 geographic areas. According to the CJR, more than 800 hospitals are required to keep the knee replacement cost below $25,565 for each episode, and they will face a financial penalty for exceeding this threshold (Japsen, 2016). According to (Barnett et al., 2019), after two years since it was implemented, the CJR model achieved a cost reduction of $812 per episode, which is a reduction of 3.1%. Motivated by the pressing needs for cost reduction and better care, we aim to optimize the clinical pathway of knee replacements, by leveraging the AI technology to learn the best clinical decisions from data.

1.2. Our Approach

Our starting point is a data set consisting full episodes of claims records from 212 successful past treatments. The records are collected from 37 care providers who are all experts in treating knee replacements. We view the clinical pathway optimization as an *intelligence crowdsourcing* problem. Care providers may have different strengths - some of them might make better decisions at certain stages of the process. Our goal is to "imitate" the best expert at different "states" of the process.

We develop a reinforcement learning-based approach to learn the optimal policy from episodic claims. Reinforcement learning is a branch of artificial intelligence (Sutton et al., 1998). It is well-suited for the sequential, uncertain and variable nature of clinical decision making. By modeling the knee replacement episode as a Markov decision process, we develop a reinforcement learning pipeline to find the optimal treatment decision at every possible state and takes its long-term effect into account. Our main contributions are as follows:

- We develop a reinforcement learning-based pipeline to analyze episodic claims records and synthesize expertise of multiple physicians into an optimal treatment policy. The pipeline provides forecast of future pathway at any intermediate stage under the optimal policy.
- The pipeline employs a dimension reduction technique to automatically aggregate histories and diagnosis in a way to maximally preserve the quality of the solution. It also uses cross validation to find the most robust solution even if data is limited, as well as to provide confidence bound of the predicted performance.
- The optimized policy reduces the average cost per episode by $1,000, which is approximately 7%. For the portion of the cost above the repayment threshold, the optimized policy reduces the average excessive premium by $500, which is approximately 33%.

If successfully validated and implemented, the optimized treatment policy is projected to result in 700 millions in savings each year in the US. The proposed approach provides a prototype for machine learning-based clinical pathway optimization. We believe it will apply to a broader scope of treatments and can significantly improve the outcome of care, leading to cost reduction and better patient experience.

2. Related Work

Reinforcement learning (RL) is an area of machine learning for solving sequential decision-making problems from data or experiment (Sutton et al., 1998). RL has achieved phenomenal successes in many fields, for examples (Mnih et al., 2015; Silver et al., 2017; Li et al., 2010; Sutton et al., 2012; Li, 2017). Compared with conventional machine learning, RL focuses on learning and acting in a dynamic environment, which is powerful to model and recommend sequential decisions to be made dynamically in clinical treatments.

There have been several attempts to apply RL in healthcare scenarios. For sepsis treatment optimization, (Raghu et al., 2017a;b) proposed a continuous-state framework to generate optimal treatment policies with deep reinforcement learning, and (Peng et al., 2018) developed a mixture-of-experts framework by combining neighbor-based algorithms and deep RL. (Nemati et al., 2016) modeled heparin dosing as a Partially Observed Markov Decision Process (POMDP), and designed a deep RL algorithm to learn individualized heparin dosing policy. (Prasad et al., 2017) used off-policy reinforcement learning to improve the dosing policy in ICU. (Parbhoo et al., 2017) used the kernel-based regression and dynamic programming to improve therapy selection for treating HIV. In these applications, the decision is often the dosage, which is a single-dimensional variable, or the decision is restricted to a small number of choices to reduce the problem's complexity.

For knee replacement treatment, there exist studies where machine learning approaches are used for prediction or classification; see (Huber et al., 2019; Navarro et al., 2018; Cabitza et al., 2018). However, these studies did not consider the optimization of clinical decision process.

Our work appears to be the first one to systematically optimize the knee replacement treatment. The proposed approach uses ideas including kernel reinforcement learning, state aggregation and dimension reduction. Kernel-based reinforcement learning was proposed in 2002 by (Ormoneit & Sen, 2002). A kernel function is needed to measure the similarity between states, based on which one can approximately solve high-dimensional Markov decision problems. It is shown that kernel-based RL produces a convergent solu-

tion that also achieves statistically consistency. Their model has been studied and extended in a number of works (Jong & Stone, 2006; Kveton & Theocharous, 2013; Barreto et al., 2016). A critical component of our approach is to learn the right kernel automatically from data. We design the kernel function by using unsupervised spectral state compression and aggregation learning, which were proposed by (Zhang & Wang, 2018; Duan et al., 2018). They provide a method to estimate the optimal partition of state space from trajectoric data to maximally preserve the transition dynamics in the data set. In this way, we are able to remedy the curse of dimensionality of high-dimensional RL by leveraging features that are learned from data.

Our approach provides a novel end-to-end pipeline for clinical pathway optimization from clinical data. It allows one to learn from records of a group of medical experts to achieve intelligence crowdsourcing by using statistical dimension reduction and reinforcement learning. The outcome is a full decision policy with robustness guarantee, together with a predictive model to forecast a patient's clinical pathway beginning from any intermediate state of the process.

3. Data

We use a data set of knee replacement claims from Cedar Gate Technologies. It is a healthcare analytics company that provides prescriptive tools to help healthcare organizations and insurance companies manage their costs and risks.

The data set contains claims records from 37 physicians and 205 unique beneficiaries (patients). It contains 212 episodes, 5946 claims and 9254 entries. Each episode is a full claims record for a patient' pathway from onset to rehabilitation and recovery. Each episode involves one physician, one beneficiary, a total cost, an episode start date, an episode end date, and a sequence of claims. Figure 1 gives a snapshot of several lines in the data set. We have masked the sensitive information to protect the patient's identity.

Attributes: Each entry contains the following attributes: episode id, beneficiary id, episode start date, episode end date, episode total cost, physician id, claim id, claim start date, claim end date, claim cost, procedure code, procedure category, diagnosis code, diagnosis category.

Date: Each episode has one "episode start date" and one "episode end date", and each claim has one "claim start date" and one "claim end date". Most claims start and end on the same day. But still, there exists some claims whose "claim start date" and "claim end date" are different.

Claim: Each claims consists of diagnoses, procedures, a start date, an end date and a cost of the claim. There are 597 different procedures, 81 procedure categories, 542 different diagnoses, and 138 diagnosis categories in this data set.

episode id	beneficiary id	episode start date	episode end date	episode total cost	physician id	claim id
[illegible]	[illegible]	[illegible]	[illegible]	18012.5	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	18012.5	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	18012.5	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	18012.5	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	18012.5	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	18012.5	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	18012.5	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	18012.5	[illegible]	[illegible]

claim start date	claim end date	claim cost	procedure code	procedure category	diagnosis code	diagnosis category
[illegible]	[illegible]	287.16	1402	232	71596	203
[illegible]	[illegible]	1160.14	27447	152	71516	203
[illegible]	[illegible]	7.19	73560	226	V5481	257
[illegible]	[illegible]	185.62	27447	152	71516	203
[illegible]	[illegible]	12014.14	NA	NA	71536	203
[illegible]	[illegible]	93.62	E0163	243	NA	NA
[illegible]	[illegible]	42.19	E0143	243	NA	NA
[illegible]	[illegible]	2611.40	1AGP1	NA	V571	254

Figure 1. First eight lines of the knee replacement data.

Almost all claims involves only one diagnosis and/or only one procedure. For example, see line 6 in Figure 1, where "NA" means the corresponding procedure or diagnosis does not exist. Some claims have more than one entries.

Beneficiaries and Physicians: There are 37 physicians and 205 beneficiaries in this data set. Each beneficiary is attributed to a unique physician.

Duration and Cost: The duration of episodes are all around 100 days, because knee replacement is a standardized and well-scheduled process. The number of claims per episode ranges from 8 to 56.

For confidentiality purposes, we are not allowed to disclose additional information about the data set.

4. A Pipeline for Intelligence Crowdsourcing from Episodic Claims

In this section, we describe the sequential decision model of the knee replacement treatment and present our methods.

4.1. Model and Formulation

We model the clinical decision making process of knee replacement as an infinite-horizon Markov decision process (MDP) with an absorbing terminal state. MDP models a dynamic state-transition process, where the state s_t at time t evolves to a future state s_{t+1} according to a transition law under the intervention of a decision maker. An instance of MDP can be described by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, C)$, where $\mathcal{S}$ is set of all possible states, $\mathcal{A}$ is set of all possible actions, P is a law of transition which is not explicitly known to the decision maker, and $C : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a state-transition reward function.

In the case of knee replacement, we model each claim as a time step. The knee replacement process may last for an indefinitely number of time steps. We model "recovery" as an absorbing terminal state, at which no future transition

or cost will be generated. According to our data set, all episodes terminate at the recovery state. In a given episode, ideally speaking, a state s_t is a collection of claims up to the time step t, and an action a_t is picked from all possible prescriptions. At time t, a physician examines the current state $s_t \in \mathcal{S}$ of a patient, chooses an action $a_t \in \mathcal{A}$ according to his/her own expertise and then the patient moves to the next state s_{t+1} according to some probability $P(s_t, a_t, s_{t+1})$. Each claim may generate a cost $C(s_t, a_t)$.

The treatment policy is modeled as a function that maps a state (claims history) to an action (a prescription), which we denote by $\pi : \mathcal{S} \to \mathcal{A}$. Our goal is to find an optimal policy to minimize the expected cumulated cost per episode, i.e., to solve the optimization problem

$$\min_{\pi} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} C(s_t, a_t) \right],$$

where $\mathbb{E}^{\pi}$ is taken over the entire pathway $(s_1, a_1, s_2, a_2, \ldots)$ on which actions are chosen according to π. Solving this ideal optimization problem precisely is intractable because the transition probabilities $\{P(s, a, s')\}$ are unknown. Given 138 different diagnosis categories and around 30 claims per episode, the number of possible states $|\mathcal{S}|$ can be as large as $138^{30} \approx 10^{64}$. The number of possible actions is 597 according to the data set. Such huge numbers of states and actions make reinforcement learning intractable.

Therefore we reformulate the policy optimization problem in the spirit of intelligence crowdsourcing. We model the state as

$$s_t = (\text{most recent diagnosis up to time } t, \text{ number of times the patient has been inpatient up to time } t),$$

which is still sufficient to capture most of the information needed for making decisions. This still yields a large state space with about 600 possible states. Further dimension reduction is needed (to be discussed in Section 5). We split the physicians randomly into multiple groups $1, \ldots, J$ and make sure that these groups have identical average cost per episode. Suppose the current state is s_t, we model the action $a_t \in \mathcal{A} := \{1, \ldots, J\}$ to be: First pick a group of physicians from $1, \ldots, J$, then pick one doctor from the group uniformly at random, finally imitate the prescription of this physician at the current state s_t. Thus the resulting state-to-prescription policy is a randomized policy. Since we have reduced $\mathcal{A}$ to be a small finite set, the action complexity has been reduced significantly. More importantly, this model can maximally leverage the physicians' expertise from their past experiences.

4.2. State Compression and Kernel Construction

To find a robust policy from limited noisy data, we need dimension reduction tools to identify latent features of the knee replacement process and further reduce the model complexity.

We employ an unsupervised spectral state compression method, which were proposed by (Zhang & Wang, 2018; Duan et al., 2018). The goal is to estimate leading features of a state-transition process $(X_1, X_2, \ldots, X_t)$ from trajectorical data. It is motivated by common latent structures of state-transition system, such that the transition kernel often admits a low-rank decomposition:

$$P(X' \mid X) \approx \sum_{z} P(X' \mid Z = z) P(Z = z \mid X).$$

The low-rank nature makes it possible to estimate the left and right singular space of the unknown transition matrix P from data.

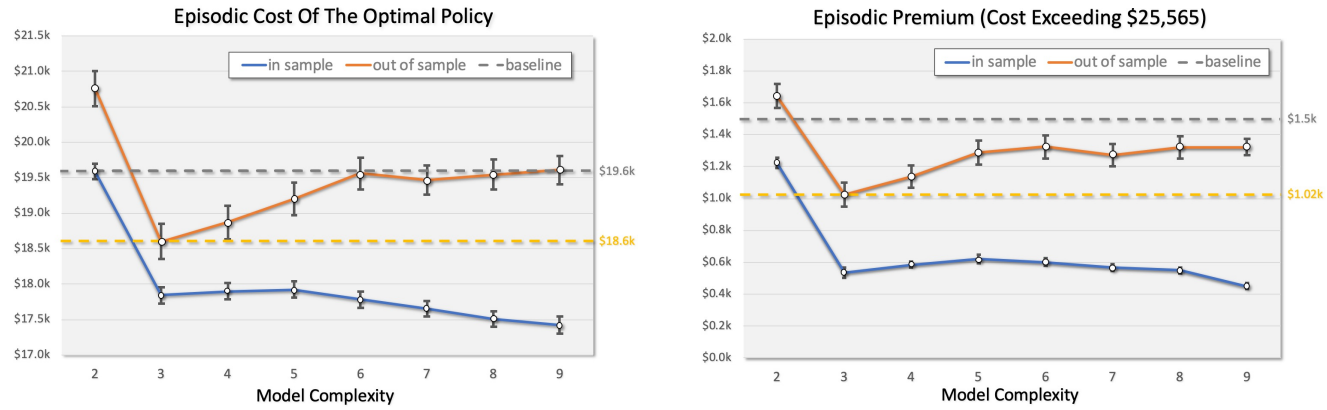
Our first step is to compute the empirical transition frequency matrix $F \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ from the data set. We order all the claims according to claim id in each episode. Every two consecutive claims gives a state-transition pair (s, s'). We compute the entry $F_{ss'}$ to be the frequency for (s', s) to appear in the entire data set. After obtaining F, we normalize it into a empirical transition matrix $\tilde{P}$, such that each row is nonnegative and sums to 1. Then we apply SVD to the transition matrix $\tilde{P} = U \Sigma V$ and get the right top k singular vectors of $\tilde{P}$, where k is a tuning parameter. It gives a matrix $\hat{V} \in \mathbb{R}^{|\mathcal{S}| \times k}$. According to (Zhang & Wang, 2018), $\hat{V}$ are referred to as Markov features that are representative of the leading structures of the unknown transition matrix.

Following (Zhang & Wang, 2018; Duan et al., 2018), we compute the optimal partition of state space $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \ldots \cup \mathcal{S}_k$ such that $P(\cdot \mid s) \approx P(\cdot \mid s')$ if $s, s' \in \mathcal{S}_j$ for some $j \in [k]$. In this way, we can aggregates states into clusters and find one best action for all states belonging to the same cluster. We compute the partition by solving the optimization problem and using the estimated $\hat{V}$:

$$\min_{\mathcal{S}_1, \ldots, \mathcal{S}_k} \min_{v_1, v_2, \ldots, v_k \in \mathbb{R}^k} \sum_{m=1}^{k} \sum_{s \in \mathcal{S}_m} ||(\hat{V})_{[s,:]} - v_m||_2^2.$$

In the experiment, we solve the preceding problem by applying the k-means method with 100 random initializations and choosing the best result.

Now that we have estimated the Markov features $\hat{V}$ and the optimal partition by using spectral state compression. We can define a kernel function for states by $K(s_1, s_2) = \hat{V}_{[s_1,:]}^T \hat{V}_{[s_2,:]}$. Alternatively, we use the estimated partition and define the kernel function by $K(s_1, s_2) = 1$ if s_1, s_2 belongs to the same block and $K(s_1, s_2) = 0$ otherwise.



(a) (b)

Figure 2. **In-sample vs out-of-sample performance of the learned optimal policy.** (a) This figure illustrates the in-sample and out-of-sample performances of the optimized policy with varying degrees of model complexity, as well as their 95% confidence intervals. The baseline (grey) is the average episodic cost in the raw data set. These curves suggest that the model tends to overfit the data as $k > 3$. The optimized policy (yellow) is projected to achieve a cost reduction of \$1,000 per episode. (b) This figure illustrates the in-sample and out-of-sample performances of excessive cost premium per episode of the best trained policy. The baseline (grey) is the average episodic premium in the raw data set. The yellow line is the best out-of-sample performance, which is achieved when the model complexity is $k = 3$. In this case, the optimized policy is projected to reduce the excessive premium per episode by approximately \$500 or 33% on average.

4.3. Kernel-Based Model Estimation

So far we have obtained a kernel function $K(\cdot,\cdot)$ that captures the similarity between any two states. It is a non-negative mapping defined as $K : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$. Now we can use the kernel function to construct an empirical MDP model $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \widehat{P}, \widehat{C})$ from the data, where $\mathcal{S}, \mathcal{A}$ are the same state space and action space as in the original MDP $\mathcal{M}$.

Let $\mathcal{D} = \{(s_m, a_m, c_m, s'_m) | m = 1, 2, \ldots, n\}$ be the set of sample transition quadruples (state, action, cost, next state). Each sample comes from two consecutive claims, and n is the total number of such samples. Now we estimate the state-transition probability matrix $\widehat{P}$ as

$$\widehat{P}(s, a, s') = \frac{\sum_{\{m|a_m=a\}} K(s, s_m) K(s', s'_m)}{\sum_{\{m|a_m=a\}} K(s, s_m)} \tag{1}$$

for any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, as long as s is not the terminal recovery state. If s is the terminal state, we always let $\widehat{P}(s, a, s) = 1$ and $\widehat{P}(s, a, s') = 0$ for all $a \in \mathcal{A}$ and $s' \neq s$. We estimate the cost function $\widehat{C}$ using

$$\widehat{C}(s, a) = \frac{\sum_{\{m|a_m=a\}} K(s, s_m) c_m}{\sum_{\{m|a_m=a\}} K(s, s_m)} \tag{2}$$

for any $s \in \mathcal{S}$ that is not the terminal state and any $a \in \mathcal{A}$. For the terminal state, we let $\widehat{C}(s, a) = 0$ for all a.

4.4. Computing the Optimal Policy

By using the appropriate kernel function, we have computed the empirical transition model (1) and empirical reward function (2) from claims data. We have obtained an empirical instance of $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \widehat{P}, \widehat{C})$, which is an approximation to the unknown true MDP $\mathcal{M}$.

Finally, we solve this empirical MDP problem $\widehat{\mathcal{M}}$ by using the value iteration method (Bertsekas et al., 1995). The value iteration makes the following iterative updates on a value vector $v \in \mathbb{R}^{|\mathcal{S}|}$:

$$v(s) \leftarrow \max_{a \in \mathcal{A}} \left[\widehat{C}(s, a) + \sum_{s' \in \mathcal{S}} \widehat{P}(s, a, s') \cdot v(s) \right], \quad \forall s \in \mathcal{S},$$

with the terminal condition $v(\text{terminal}) = 0$. The value iteration is known to provide a sequence of solutions converging to the optimal value function $v^* \in \mathbb{R}^{|\mathcal{S}|}$. Then we obtain the optimal policy π^*:

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \left[\widehat{C}(s, a) + \sum_{s' \in \mathcal{S}} \widehat{P}(s, a, s') \cdot v^*(s) \right].$$

The computed π^* prescribes a physician group to follow for each state. The final treatment policy is a set of conditional distributions over prescriptions, which are estimated from the data set following π^*.

5. Empirical Results

In this section, we present our main results. Empirical simulation shows that the proposed pipeline produces a robust policy and results in significant cost reduction.

5.1. Projected Performances and Cross Validation

We divide the diagnosis categories into k clusters by using the spectral state clustering method in Section 4, for $k = 2, \ldots, 9$. The partition map is further used to build a kernel function for the state space $\mathcal{S}$. We view k, the rank parameter in state compression, as a measure of the model complexity. It is a tuning parameter that controls the tradeoff between model reduction error and overfitting error.

We conduct both in-sample and out-of-sample experiments. In in-sample experiments, we train the treatment policy to minimize cost over the entire data set, and test its performance on the same raw data. In out-of-sample experiments, we use 4-fold cross validation. We split the raw data set into 4 equal-size segments randomly, and compute the optimal policy using three out of the four segments and test its performance on the remaining data. Each experiment is repeated randomly for 2,000 times and for all values of k. The results are given in Figure 2. In particular, Figure 2 (a) shows the average cost per episode obtained from in-sample and out-of-sample experiments across values of k.

Note that according to (Japsen, 2016), hospitals have to keep their costs below \$25,565 or they will face a financial charge. We define the episodic excessive cost premium as

$$\text{episodic premium} = \max\{0, \text{episode cost} - 25565\}.$$

In Figure 2 (b) we compare the episodic premium from the in-sample and out-of-sample experiments.

The results of Figure 2 are consistent with typical in-sample vs out-of-sample error curves for training machine learning models. It is expected that the in-sample error decreases monotonically as the model complexity k increases, however, over-fitting would occur when k is too large. The out-of-sample error curves show that over-fitting occurs when $k > 3$, therefore we pick $k = 3$ to be the best model complexity for this data set. We take the out-of-sample performance at $k = 3$ as the predicted performance of the optimized policy. In this case, the average cost per episode reduces by 7 percent and the average cost premium per episode reduces by 33 percent.

5.2. Cost Distribution and Tail Improvement

In Figure 3 we compare the histograms of episodic cost over the raw data and after optimization. The after-optimization cost distribution is obtained by simulation. The percentages of episodes costing over \$30k are reduced significantly. Such a reduction mitigates the financial risk faced by healthcare providers, as they are penalized for costs over \$25,565 according to CMS.

The tail improvement is significant for reasons beyond financial risk. It is well known that excessive medical cost in medical treatment is highly correlated to unsuccessful cases, for example post-surgery complications often lead to extra surgeries, higher costs and traumatic discomforts. Therefore the tail cost reduction suggested by Figure 3 suggests that the optimized clinical policy may reduce unnecessary medical complications and extra inpatient surgeries. Therefore the proposed pipeline not only reduces financial costs but also improves the quality of care.

5.3. Clinical Pathway Forecast Using The Optimal Policy

Now we have estimated the best clinical policy from data. Our model allows us to predict the entire clinical pathway of a typical knee replacement treatment. We simulate the knee replacement process under the optimized policy based on existing claim episodes. This generates 10,000 sample trajectories. Then based on the sample trajectories, we compute the empirical distribution of diagnosis categories per each day. Figure 4 (a) illustrates the distribution over states as time progresses, when the optimal policy is used.

Our model can also forecast the future clinical pathway of a patient from any given time and given state during the knee replacement episode. Figure 4 (b)(c)(d) illustrates the forecasted pathway from day 25, day 50, day 75, respectively, conditioned on the event that the current diagnosis belongs to the category of osteoarthritis. Figure (e) illustrated the forecasted pathway from day 50, conditioned on the event that the current diagnosis belongs to the category of non-traumatic joint disorders. We see that the model is able to predict the conditional distributions of future diagnosis and treatments given any particular time and state of the treatment. These forecasts can make physicians better informed while making clinical decisions.

6. Conclusion

In this paper, we propose a reinforcement learning-based pipeline to optimize the knee replacement treatment from medical claims. It produces a clinical policy via intelligence crowdsourcing from experts. Empirical validation suggests the optimized policy reduces the average cost per episode by \$1,000, which implies a potential savings of 700 millions each year in the US. We believe this result gives a proof-of-concept that reinforcement learning can significantly improve clinical decisions. For future work, we will focus on personalizing the knee replacement treatment by using data sets of large scales and higher-resolution models.

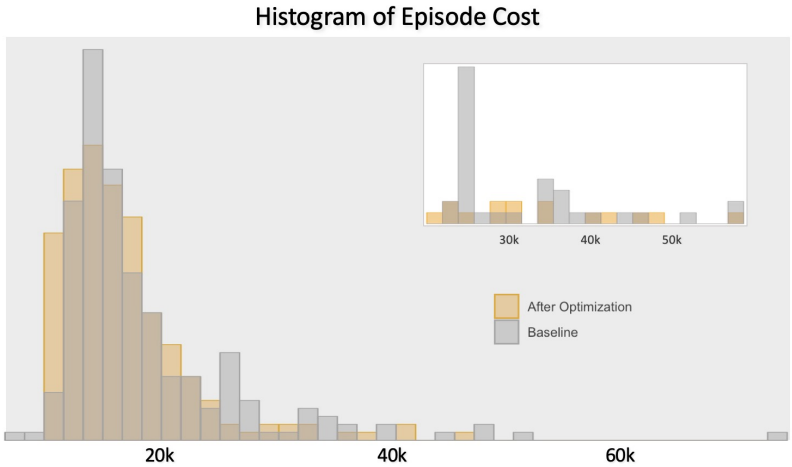


Figure 3. **Histogram of costs per episode.** This plot gives the distribution histogram of cost per episode in the raw data (grey) and the simulated cost histogram if the optimized policy is used (yellow). The baseline cost per episode is on average $19,559 with standard deviation $8,101, while the improved episode cost after optimization is on average $17,781 with standard deviation $5,750. In the upper right corner, we zoom in the tail distribution and examine the costs above $25,565. The excessive cost premium is $1,500 in the raw data, which reduces to $566 after optimization.

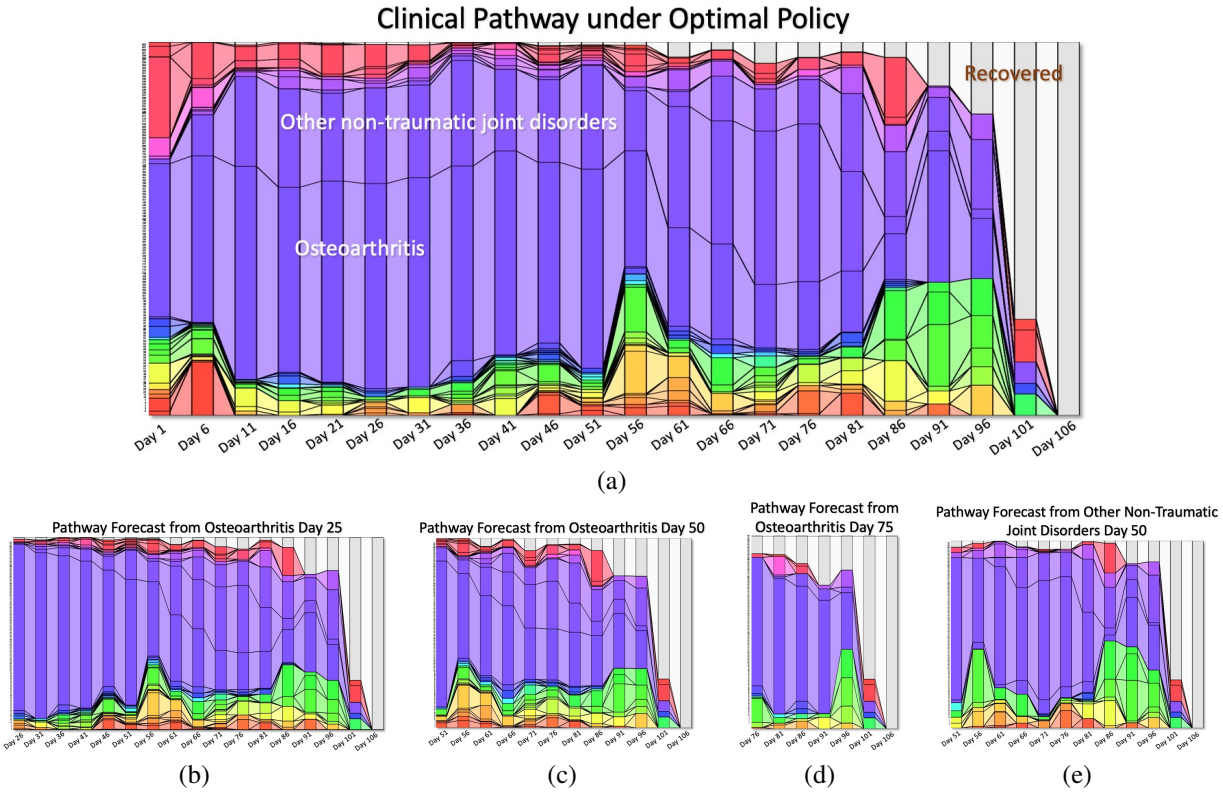


Figure 4. **Clinical pathway forecasts under the optimized treatment policy.** X-axis corresponds to the day of treatment, and Y-axis corresponds to 138 diagnosis categories (each has a unique color). For each day, the bar plot shows the probability distribution of a patient's state over the 138 diagnosis categories. Figure (a) gives the full clinical pathway for a typical patient. Figure (b), (c), (d) give forecasts of the future pathway from day 25, day 50, day 75 respectively, conditioned on the event that the current diagnosis belongs to the osteoarthritis category. Figure (e) illustrated the forecasted pathway from day 50, conditioned on the event that the current diagnosis belongs to the category of non-traumatic joint disorders.

References

Barnett, M. L., Wilcock, A., McWilliams, J. M., Epstein, A. M., Joynt Maddox, K. E., Orav, E. J., Grabowski, D. C., and Mehrotra, A. Two-year evaluation of mandatory bundled payments for joint replacement. *New England Journal of Medicine*, 380(3):252–262, 2019.

Barreto, A., Precup, D., and Pineau, J. Practical kernel-based reinforcement learning. *The Journal of Machine Learning Research*, 17(1):2372–2441, 2016.

Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P., and Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Cabitza, F., Locoro, A., and Banfi, G. Machine learning in orthopedics: a literature review. *Frontiers in bioengineering and biotechnology*, 6, 2018.

CMS. Comprehensive care for joint replacement model. https://innovation.cms.gov/initiatives/CJR, 2016.

Doorhof, D. Using reinforcement learning to improve clinical decision making in neonatal care. Master's thesis, 2018.

Duan, Y., Ke, Z. T., and Wang, M. State aggregation learning from markov transition data. *arXiv preprint arXiv:1811.02619*, 2018.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

Huber, M., Kurz, C., and Leidl, R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC medical informatics and decision making*, 19(1):3, 2019.

Japsen, B. Medicare's bundled fees hit knee, hip replacements, Apr 2016. URL https://www.forbes.com/sites/brucejapsen/2016/04/01/starting-today-medicares-bundled-fees-hit-hospitals-hard/#607be81926d9.

Jong, N. and Stone, P. Kernel-based models for reinforcement learning. In *ICML Workshop on Kernel Machines and Reinforcement Learning*. Citeseer, 2006.

Kveton, B. and Theocharous, G. Structured kernel-based reinforcement learning. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670. ACM, 2010.

Li, Y. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

Luo, J., Wu, M., Gopukumar, D., and Zhao, Y. Big data application in biomedical research and health care: a literature review. *Biomedical informatics insights*, 8:BII–S31559, 2016.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, 2014.

Martin, G. M. Patient education: Total knee replacement (arthroplasty)(beyond the basics). *UpToDate. com*, 2017.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2017.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529, 2015.

Navarro, S. M., Wang, E. Y., Haeberle, H. S., Mont, M. A., Krebs, V. E., Patterson, B. M., and Ramkumar, P. N. Machine learning and primary total knee arthroplasty: patient forecasting for a patient-specific payment model. *The Journal of arthroplasty*, 33(12):3617–3623, 2018.

Nemati, S., Ghassemi, M. M., and Clifford, G. D. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2978–2981. IEEE, 2016.

Olsen, L., Saunders, R. S., Yong, P. L., et al. *The healthcare imperative: Lowering costs and improving outcomes: Workshop series summary*. National Academies Press, 2010.

Ormoneit, D. and Sen, Ś. Kernel-based reinforcement learning. *Machine learning*, 49(2-3):161–178, 2002.

Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V., and Doshi-Velez, F. Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Lehman, L.-w. H., Ross, A., Faisal, A., and Doshi-Velez, F. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, pp. 887. American Medical Informatics Association, 2018.

Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., and Engelhardt, B. E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017a.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. *arXiv preprint arXiv:1705.08422*, 2017b.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2018.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Sutton, R. S., Barto, A. G., et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Sutton, R. S., Szepesvári, C., Geramifard, A., and Bowling, M. P. Dyna-style planning with linear function approximation and prioritized sweeping. *arXiv preprint arXiv:1206.3285*, 2012.

Zhang, A. and Wang, M. Spectral state compression of markov processes. *arXiv preprint arXiv:1802.02920*, 2018.