Evaluating the effectiveness of a reinforcement learning model in customizing colonoscopy screening policies

DANIËL BLIK, University of Twente, The Netherlands

Colorectal and rectum cancer (CRC) is a major global health concern, contributing to both morbidity and mortality rates and decreasing the quality of life of patients. This cancer originates from small growths also referred to as polyps. These polyps if not detected and removed promptly can grow into tumors. To detect these polyps a procedure called colonoscopy is used. This procedure allows for the visual examination of the colon's interior and the removal of polyps. The colonoscopy procedure is one of the most effective procedures for the detection and removal of polyps. However, the optimal timing policies for colonoscopy procedures remain a topic of debate. This thesis aims to contribute to the ongoing discussion on colonoscopy screening guidelines. This paper will evaluate existing screening guidelines and existing reinforcement learning models by literature review. After which it will evaluate the performance of a new model based on partial observability utilizing a simulation. This model takes into account more personalised aspects of the policy recommendation. Finally, the paper attempts to make a valuable conclusion which may contribute to the ongoing debate on colonoscopy screening guidelines.

Additional Key Words and Phrases: Colonoscopy, partially observable Markov decision process, Reinforcement learning, QALY

1 INTRODUCTION

Colorectal and rectum cancer (CRC) is a form of cancer contributing to both morbidity and mortality rates on an international level[3, 12]. This form of cancer stems from benign growths known as polyps, which have the ability to develop into malignant tumours if not detected and removed promptly. CRC is the third most common type of non-skin cancer and is the second leading cause of cancer death in the United States. In 2021, an estimated 149,500 people in the United States were diagnosed with CRC, and 52,980 people died from it[3]. In the EU approximately 170,000 people die from CRC annually with an expected rise [12]

The primary method for early detection and prevention of CRC is through colonoscopy, a procedure that allows for the visual examination of the interior of the colon and the removal of polyps. Despite its effectiveness, the optimal policies for colonoscopy screening remain a topic of ongoing debate in the medical community.

During a colonoscopy, the rectum and entire colon are examined using a flexible lighted tube with a lens for viewing and a tool for removing tissue. This is often experienced as an uncomfortable experience with patients having to undergo diet and medication changes before the test, their colon needs to be cleansed and sedation is needed[3].

Currently, most cancer screening's such as colonoscopies are scheduled by the usage of broad recommendations and guidelines. In the case of colonoscopy [2] found that a majority of these guidelines recommend screening an average-risk individual between the age of 50 and 75 every 10 years. However, these guidelines do not consider any patient-specific circumstances such as increased risk profiles, age, gender etc.

However, these general guidelines for screening suffer from several logistical and cost-effectiveness drawbacks. Firstly, screening every 5 to 10 years from the age of 50 onward would imply screening whole populations at the same frequency, resulting in significant costs for the healthcare system and logistical burden. This is an issue since in the EU only 14 percent of the citizens participate in CRC formal population-based screening programs. One of the key aspects for improvement was proper capacity[12]. If the EU was able to diagnose patients at earlier stages up to 130,000 lives could be saved per year, and more than 3 billion euros in healthcare budget could be saved [12]. With the usage of personalized screening policies the capacity would be optimized as not an entire population group has to be screened at once. Secondly, broad guidelines do not address surveillance after malignant screening results. Lastly, as mentioned before patients might require shorter screening intervals or can tolerate a longer interval based on their risk profile based on for instance their genetics, age, and gender as discussed in [5].

To solve these issues personalized surveillance recommendations can be made by using reinforcement learning as discussed in [11]. To make decisions on screening policies many factors need to be taken into consideration. Firstly as the outward health state of a patient is directly observable and colonoscopy is not a perfect test in the detection of adenomatous polyps partial observability should be employed. This thesis will evaluate the model of [1] which employs a partially observable markov decision process(POMDP) as its model.

Other papers have also employed a POMDP model like [5] incorporated the personal risk of having CRC and adenomatous polyp and other factors such as age and gender into screening their partially observable Markov decision process (POMDP) model. [13] modelled the personalized breast cancer screening as a sequential decision-making process and solved it through envelope-Q learning.

The model used in [9] incorporates factors such as screening frequency, initial screening age, and partial compliance in their POMDP model, but does not incorporate other factors such as gender, race and family history. which are factors likely to affect the CRC disease chances. One trend noticeable in most papers evaluating screening policies is the usage of partial observability.

This bachelor thesis aims to evaluate existing colonoscopy policies with those generated by a reinforcement learning model. The evaluation will evaluate the effectiveness of the policies by means of a simulation where the quality-adjusted life years(QALY) values are compared between policies like what was done in the paper of [6] for patients by age and recommendation. To do this first existing policies are gathered and reviewed utilizing a literature review. After this existing reinforcement learning models and their performances are reviewed by a literature review. This should give us insight

TScIT , July 5, 2024, Enschede, The Netherlands

into the current employment of reinforcement learning models in generating cancer screening recommendations.

This should give us the answers to the following questions:

(1) How cost-effective are the current colonoscopy policies?
(2) How are reinforcement learning models used for generating cancer screening policy?

With the answers to these questions, this thesis will build up to answer the main research question underlying this thesis namely: How can reinforcement learning models contribute to the development of optimal colonoscopy screening policies?

Firstly in chapter 2 the literature review is done where both the sub-research questions are answered. This will be followed up in chapter 3 with a methods of research where it is explained how the model developed in [1] will be evaluated. After which the results of this will be shown and elaborated on in the results section in chapter 4. From the results, in chapter 5 the conclusions will be discussed in the conclusions section. After which in the last section the conclusions, results and future work will be discussed in chapter 6 with the the discussion section.

2 LITERATURE REVIEW

2.1 How effective are the current colonoscopy policies?

The paper [7] evaluates the cost-effectiveness of colorectal cancer screening. They do this by conducting a review according to the framework for reviews of economic analyses[7]. The paper reviews 32 studies to determine the cost-effectiveness of 4 screening strategies: colonoscopy on a 10-year interval, annual guaiac fecal occult blood test(FOBT), 5-yearly sigmoidoscopy, a combination of 5-yearly sigmoidoscopy and annual guaiac FOBT. on a 10-year interval. colonoscopy on a 10-year interval is the most common colonoscopy guideline used. The cost-effectiveness was expressed as a value of discounted life year gained(LYG) this was evaluated by 8 different models. From these 8 different models, from these models 5 found 10-year colonoscopy to be the most effective in discounted LYG[7]. The remaining 3 models found that the combination of sigmoidoscopy and FOBT was the most effective. The paper also included 5 other models which did not include the combination of sigmoidoscopy and FOBT, resulting in 12 studies on annual guaiac FOBT, 5-yearly sigmoidoscopy and 10-yearly colonoscopy. From these 12 studies, all 12 studies found colonoscopy to be the most effective screening [7]. Furthermore, in 8 of the 12 studies, it was found that colonoscopy was the preferred method at a willingness to pay 50,000 US dollars per LYG.

Furthermore, the effectiveness of colonoscopy screening strategies is assessed in the paper of [10]. This paper evaluates the effectiveness of different types of colonoscopy screening strategies by developing a partially observable Markov decision chain(POMDC). By using their model they evaluated a multitude of colonoscopy strategies both fixed-interval strategies and observation-based strategies. For the observation-based strategies, they used existing clinical classification. The clinical classification defines 4 distinct groups in terms of precancerous adenoma prevalence. Thus [10] evaluates the following strategies in the paper:

- 10-yearly colonoscopy (fixed-interval)
- 20-yearly colonoscopy (fixed-interval)
- group 1: 10 yearly colonoscopy (observation-based)
- group 2: between 5 to 10-yearly colonoscopy (observation-based)
- group 3 and 4: 3-yearly colonoscopy (observation-based)

Next to this the paper also takes the effects of several parameters related to CRC screening strategy design into consideration. This includes the initial age to start screening, the age to stop screening, and the screening compliance rate.

The results in the paper were analysed on cost-effectiveness using quality-adjusted life years (QALY) of the policy in comparison with the no-screening performance. The numerical studies of the paper showed that the current screening guidelines and variations of it are cost-effective compared to not doing any screening[10]. Furthermore, it was found that varying the screening interval is especially in observation group 2 very influential. It showed that a larger interval for observation group 2(e.g. more towards the 10-yearly interval) is more cost-effective[10]. Finally, the study finds that the specified parameters had significant results. For example, varying the initial screening age had a significant impact on the cumulative QALYs as initiating screening at an earlier age led to increased cumulative QALY's[10].

Considering the results from the papers of [7, 10] we can answer our first question, How effective are the current colonoscopy policies? In the reviewed papers we find that existing policies perform reasonably well in terms of cost-effectiveness. In comparison with other screening methods, we find that below the range of 50.000 US dollars per LYG colonoscopy performs better[7]. Furthermore, we find that below this threshold of 50.000 US dollars, there is little difference between different types of screening strategies.

2.2 How are reinforcement learning models used for generating cancer screening policies?

Research on the usage of reinforcement learning models for generating cancer screening policies has already been done before.

Firstly, in the paper of [9] a POMDP model was formulated to optimize the biopsy policy in prostate cancer active surveillance, with the goal of minimizing the expected delay in detecting high-risk cancer and minimizing the number of lifetime biopsies [9].

The model made a number of assumptions. Among others, It simplified the stochastic process of prostate cancer progression to a first-order Markov chain[9]. The model also assumed perfect specificity for biopsies [9]. Furthermore, it was assumed that the annual cancer progression rate was stationary and not age-dependent which was validated by previous studies [9].

The study provided insight into optimizing biopsy decisions as the POMDP model revealed structural properties that can guide model-based biopsy policies. These properties mainly have to do with the belief state used in a POMDP model for example the paper found that the solution to the POMDP model was a control-limit type policy meaning that there is a threshold on the element of the belief vector, which represents the probability of being in a certain state, where below which it is optimal to defer biopsy and above which it is preferred to conduct biopsy[9].

Secondly, In [5] a POMDP was developed to recommend colonoscopy screening. In this model, a multitude of risk factors were included mainly static risk factors which are static features of a patient such as age and dynamic risk factors which are changing features of a patient such as history of polypectomy intervention. The static risk factors used in the model were age, gender and history of CRC or adenomatous polyps. The dynamic risk factors used in the model consisted of the history of polypectomy or CRC treatment since this can influence the progression of colectoral cancer over time [5].

In order to take the previously mentioned risk factors into account the researchers define completely observable risk-levels namely low-risk, high-risk and post-CRC[5]. These risk levels allow for the creation of customized colonoscopy screening recommendations since they serve as belief states in the model. The model believes that a patient is either low-risk, high-risk or post-CRC. The belief state is updated on every decision epoch, based on the screening observations[5]. The model operates under the assumption that screening starts at age 50, where the researchers justify this since almost all guidelines suggest initiating screenings at age 50[5]. The model was solved to maximize total quality-adjusted life years(QALY) at the age of 50+N. In order to do this they utilized an Expectation-Maximization (EM) algorithm. They compared their results against the MISCAN model from [8] and concluded that the results of the POMDP model lie very close to those obtained by the MISCAN model [5]. Finally, they concluded that current guidelines while effective, could have been improved by considering factors like age, gender and personal history in screening decisions to enhance QALY scores and reduce colorectal cancer, risk and mortality.

Lastly, in [10] the cost-effectiveness of colonoscopy screening strategies are asses using a partially observable Markov chain(POMC) model. The paper considers detailed adenomatous polyp states and estimated transition probabilities in their model. Where the transition probabilities are based on longitudinal data from a specific population cohort. The paper highlights the importance of partial observability in reinforcement learning models when dealing with colonoscopy screenings since colonoscopy screenings inherently sometimes only view a partial amount of the actual existing polyps [10].

To incorporate this partial observability the paper employs a Bayesian approach to the belief state(the state believed since it cannot be fully observed) [10]. Each state in the POMC model is assigned a QALY multiplier to reflect the impact on the individual well-being. The POMC model provided enhanced accuracy and overall better QALY scores in comparison with general screening policies [10]. However, the model was trained and evaluated on a very specific dataset, the POMC model also excluded other CRC risk factors in the model which suggests areas for future improvement and research.

From the papers analyzed it is seen that most papers conclude that the usage of a reinforcement learning model has potential to improve the existing guidelines. However, most models have several limitations mainly based on a number of assumptions made in order to make the model work. For models discussing colonoscopy screening these factors are mostly related to risk factors, personal history and age. Furtehmreo, the models are mostly trained on a limited dataset.

Therefore we can answer our sub-research question with the answer. Currently, reinforcement learning models are in development and add to the ongoing discussion on cancer screening policies. However, most models are developed with several assumptions or are trained on limited datasets limiting their potential usage to highlighting important factors that need to be considered when developing cancer screening policies.

3 METHODS OF RESEARCH

To find an answer to the previously defined research question a simulation is performed to determine both the overall effectiveness and cost-effectiveness of a recent reinforcement learning(RL) model the model of [1] will be used. The simulation will simulate three policy types. Firstly, the policy generated by the RL model which is obtained through a reinforcement learning algorithm, a 10-yearly policy as this is determined to be the most used and most effective existing policy as seen in 2 and a no-screening policy to serve as a control group.

3.1 POMDP Model

The model of [1] defines a partially observable Markov decision process(POMDP) model intended to personalize the screening policies based on the personal history of CRC or polyp detection and the patient's age[1]. where the transition probabilities have been calculated based on patient colonoscopy screening data of 5 veteran army hospitals [1]. The model uses a definition of t for time epochs in years as it assumes decisions are made on an annual basis. for the information gained from the POMDP model we find the relevant information of S_T, for the state of a patient at time t, $r_t(s, a, o)$ which is the immediate rewards a patient receives when in state s taking action a and receiving observation o at time t and

3.2 Q-learning

To determine the effectiveness POMDP model Q-learning will be employed to extract possible policies from the model. Q-learning is an algorithm that determines the optimal path for an agent in a given context. It does this by employing The Q-learning algorithm which is an incremental algorithm for estimating an optimal decision strategy in an infinite-horizon decision problem[4]. As in the context of a colonoscopy a possible RL agent has to decide on whether or not to perform a colonoscopy at a given point in time. Therefore a Q-learning algorithm is suitable.

The Q-learning implementation in the simulation is based on the Bayesian principles. As the model on which the Q-learning algorithm will be applied is non-deterministic and has uncertainty the Q value rewards might become highly varied, especially in the earlier Q-learning stages. In order to avoid this a Bayesian implementation is used. The Bayesian implementation addresses this problem of high variability by maintaining a probability distribution over the Q-values rather than a single-point estimate. This probabilistic approach allows the agent to quantify and manage the uncertainty in its Q-value estimates, leading to more robust decision-making. The implementation of the algorithm can be found in A.1

3.3 Simulation

To evaluate the model against existing policies a simulation using Python is utilized. In this simulation, an n amount of patients is generated with an initial polyp state based on the model's probability of a patient being in that state. For the n amount of patients a simulation is executed 25 time periods. Because as found in 2 most existing colonoscopy guidelines start screening at the initial age of 50 and stop at the age of 75. Since we are evaluating against a 10-yearly colonoscopy screening policy the timeline of 25 years is considered. In each year for 25 iterations the action is determined by policy. In the case of a no-screening policy, the action will always be to not do any screening. For the 10-year screening policy, a screening is done every 10 years. For the reinforcement learning policy, the decision is based on the Qtable. At every time epoch, the Q-table is referenced. Here the option is chosen based on the highest q value of the action for example, if it is believed that the patient is at state 30 the highest q value at state 30 will be chosen which corresponds to either performing a colonoscopy or not doing so.

In the simulation, the next state for the patient is determined in two ways for which two simulations will be run. firstly the state transitions will be based on the transition probabilities provided by the model of [1]. However, with these transition probabilities, the general path of a patient is less aggressive and perhaps might not represent a patient with aggressive colorectal cancer. Therefore another simulation will be evaluated where the state transition is based on a random amount of increase in polyp state to also evaluate the performance under higher polyp states. This increase in the polyp state will be simulated by picking two numbers between 0 and 90 - the current state and taking the minimum from these two numbers. This will result in picking a number which will never result in a polyp state above 90 where the chances of picking a higher number are slimmer than picking a lower number.

Furthermore, for each patient simulation the amount of performed colonoscopies, total reward in QALY, states visited and the final state are collected. This results in a dataset of n entries in 3 columns representing the 3 different policies with the simulation results for n patients.

4 RESULTS

To determine the results we simulate using the previously discussed Python simulation for n = 100000 patients.

4.1 Transition probability simulation

First, the results of the transition probability simulation will be discussed. this simulation is based on the transition probability between the different patient states considered by the model in [1]. In these transitions, the probability is defined for each state. based on this probability, in each simulated year, there is a new state picked from the list. in Table 1 we highlight the general performance metrics by average reward, average end state and average amount of screenings. In the results, we discuss QALY and evaluate the performance by using screenings per QALY as the costs of a colonoscopy screening vary between regions in the world.

Table 1. Performance metrics of transition probability simulation

Metric	Reinforcement Learning	10-yearly	no screening
Average reward	16.73	18.0	16.4
Average end state	28.31	11.59	39.85
Average amount of screenings	0.92	2.00	0.00
QALY per screening	18.18	9.0	-

As seen in Table 1 we see that the 10-yearly colonoscopy screening policy is most effective on a pure reward basis. The 10-yearly policy has a 1.27 higher QALY score, but has 1.08 more screenings on average. Resulting in a much worse score for QALY per screening. The results of which have been compiled into tables highlighting the average amount of screening, average terminal reward and average end state. Furthermore to provide insight into the data a line graph of the path by policy has been generated as well as a box plot.

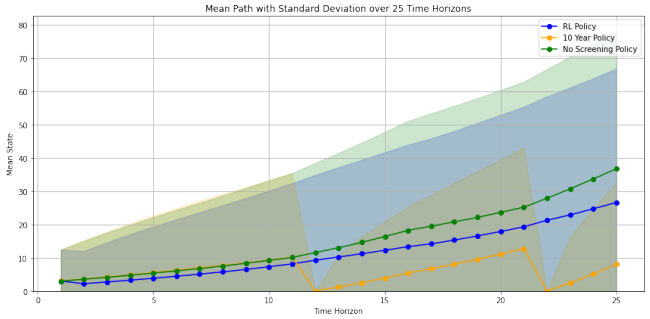

Fig. 1. Plot of mean path including standard deviation for transition probability simulation

In Figure 1 we see that all paths remain relatively the same for the first 10 years. After which the paths separate. Something seen from this path plot is the way the 10-yearly policy behaves. Where at each 10 yearly a sharp drop can be seen in the mean state of the path. This may highlight the reliance of the policy on a well-executed colonoscopy screening. After the steep drop, we immediately notice a sharp incline, meaning that if a colonoscopy screening were not to be as effective this would result in a very sharp rise in a patient's state without the initial drop. We notice that the reinforcement

learning(RL) policy tends to separate more from the no-screening policy path as time goes on. This might indicate more stability in the RL policy.

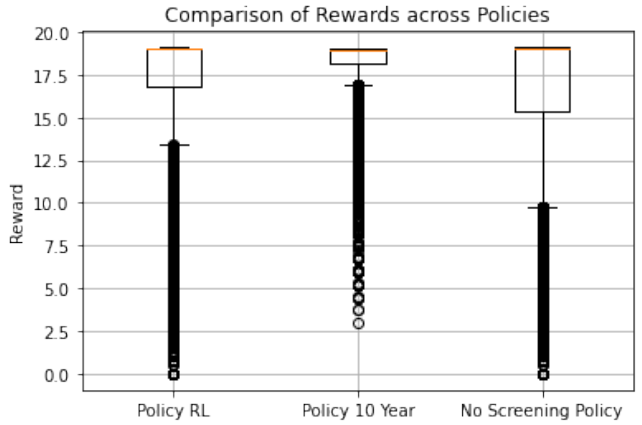

Fig. 2. Box-plot of rewards by policy for transition probability simulation

In the box plot in Figure 2 we see that the no-screening policy offers the most extreme values as the box and thus the interquartile range(IQR) is the largest of them all. We also see that the RL policy produces a larger box than the 10-yearly policy does. Potentially highlighting that the 10-yearly policy is slightly more reliant than the RL policy is.

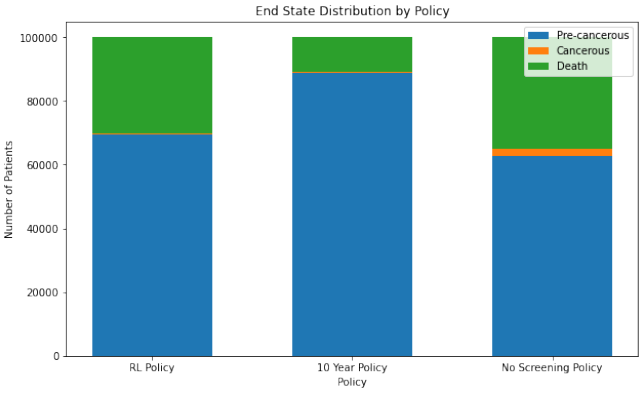

Fig. 3. End state distribution by policy for transition probability simulation

Figure 3 highlights the end states by policy here we can see that the 10-yearly policy has a smaller amount of patients in the death state. Confirming the fact that generally speaking the 10-yearly policy performs better with regards to reducing the amount of patients in either death or cancerous states. Interestingly in this simulation, very few patients reach and stay in a cancerous state. This is unexpected since in the no-screening colonoscopy some patients do stay in the cancerous state.

4.2 Random state transition simulation

In the random state transition simulation, we consider a more aggressive and random transition between states. In each year a new state is picked based on a random probability. As previously elaborated on a random number between 0 and 90(the death state) is generated where the larger the number gets the smaller the probability that this number is chosen. This provides a much more aggressive and less predictable patient path than the transition simulation and should provide us with different simulation data.

Table 2. Performance metrics of random state transition simulation

Metric	Reinforcement Learning	10-yearly	no screening
Average reward	14.49	12.89	4.51
Average end state	56.83	71.87	89.98
Average amount of screenings	5.12	2.00	0.00
QALY per screening	2.83	6.45	-

In the performance Table 2 we see that in this instance with a more aggressive and random state transition that average reward and end state the RL policy is a more effective policy than the 10-yearly policy. However, in this simulation, the RL policy was outperformed by the 10-yearly policy on screenings per QALY. This perhaps demonstrates that on a more aggressive and random approach, the RL policy recommends more screenings to combat an aggressive polyp transition whilst sacrificing its cost-effectiveness.

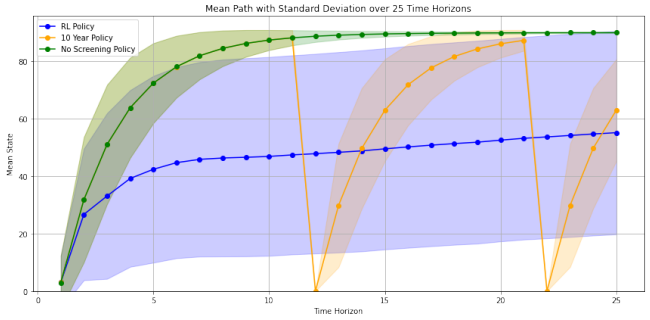

Fig. 4. Plot of mean path including standard deviation for random state transition simulation

In Figure 4 we see the mean path generated by the different policies under the random state transition simulations. In this graph, it can be seen that this is a more unpredictable and aggressive state transition resulting in very sharp rising graphs. However, we do see that the RL policy path stays relatively the same as the path displayed in Figure 1 of the probability transition simulation. Where despite the initial difference in sharp rise the path remains the same from time years 5 onward. Indicating that indeed the RL policy sacrifices some cost-effectiveness to compensate for the aggressive polyp development in the patient.

From the boxplot in Figure 5 we can conclude that despite having a lot of extreme outliers the box is positioned much higher, however, the inter-quartile range is still slightly larger compared to the 10-yearly policy. Indicating that despite the RL providing higher rewards it is ever so slightly more variable.

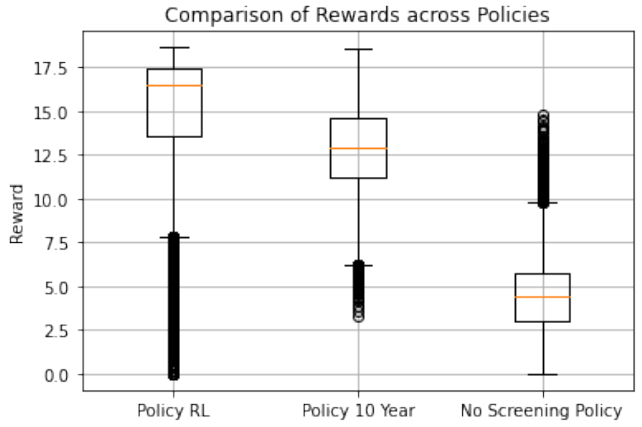

Fig. 5. Box-plot of rewards by policy for random state transition simulation

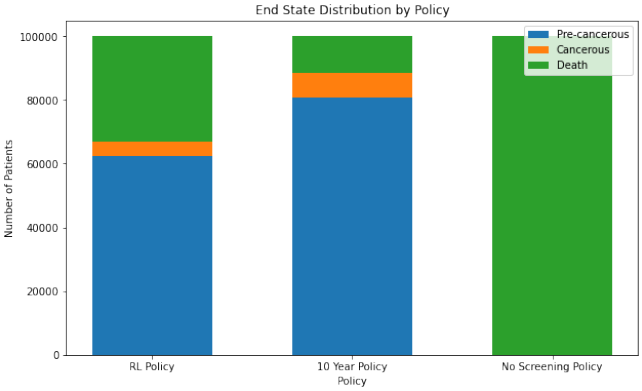

Fig. 6. End state distribution by policy for random state transition simulation

Figure 6 depicts the end states by the policy. In this graph, it can be seen that the no-screening policy always ends up in the death state. Confirming that the random transition depicts a very high-risk colorectal cancer. Between the RL policy and the 10-yearly policy, we notice that significantly fewer patients end up in the death state under the 10-yearly policy, however in the 10-yearly policy more patients end up in a cancerous state than in the RL policy.

5 CONCLUSION

This study evaluated the effectiveness of a reinforcement learning model in comparison to existing colonoscopy policies. The evaluation was done using a Python simulation which simulated results for 100000 patients. Two scenarios were simulated one scenario for a state transition based on the probabilities of the reinforcement learning model and a scenario simulating for random CRC progression.

Firstly, for the first simulation using the probability transitions from the reinforcement learning model it can be concluded that the RL policy yields a slightly lower QALY reward than the existing 10-yearly policy, but still yields a higher reward than the control of the no-screening policy. Despite the lower QALY result from the RL policy it only employs 0.98 screenings on average resulting in significantly higher screenings per QALY than the 10-yearly policy with a score of 18.18 in comparison with 9.0. Additionally from the box-plot in Figure 2 it can be seen that the RL policy has a larger variability than the 10-yearly policy indicating that for some patients the employment of RL might be a better option as less reward is sacrificed while using a lower amount of screenings. In the box plot, it can also be seen that the 10-yearly policy has a number of outliers which might have a better performance under the RL policy.

Secondly, in the second simulation using the random state transitions we see the result flipped. It is seen that the reward of the RL policy is significantly higher than the 10-yearly policy but the RL policy employs many more screenings than before as in this simulation it employed on average 5.12 screenings. Therefore in this simulation, the 10-yearly policy is actually more cost effective at 6.45 QALY per screening.

Lastly, from the box plots we see that in both simulations the RL policy is a bit more variable than the 10-yearly policy indicating that for some patients and possibly for some patient groups the RL policy is much more effective than for others in comparison with the general 10-yearly policy which has a lesser variability and thus might be better applicable to a larger patient group.

In conclusion, it was found that simulated for the regular patient transition the RL policy yields much better cost-effectiveness than the general 10-yearly policy whilst only resulting in 1.27 less QALY. indicating that for a general patient population the usage of the RL policy improves the cost-effectiveness of the colonospy screenings. However, given a more aggressive cancer the RL policy is less cost-effective but yields a significantly better average QALY for a patient than the general 10-year population does suggesting the RL policy is better employed for a high-risk individual when considering the best QALy outcome for a patient.

Thus, we find the answer to our main research question to be: given proper socio-economical implementation, reinforcement learning models can significantly improve the cost-effectiveness of colonoscopy screenings in comparison with existing policies. For patients with a high-risk profile, the usage of RL-based policy can provide a more effective screening recommendation than general screening policies can but in a less cost-effective manner.

6 DISCUSSION

6.1 simulations results

The first simulation which is based on the transition probabilities obtained from the model of [1] provides some interesting outcomes. Firstly, the amount of outliers found in the box-plot is somewhat surprising. For both the RL-policy as the 10-yearly policy a large number of outliers is found. This may influence the results of the simulation. The large number of low-reward outliers might have affected the mean results on which the conclusion are largely based. In this simulation also a surprisingly small amount of patients ended up in the cancerous state, most patients either ended in a pre-cancerous or death state but practically zero patients ended up in the cancerous state. The reason for this and the outliers might lie in the dataset used. The transition probabilities used are compiled by the RL model

of [1] based on a dataset of 1400 patients from 5 different VA hospitals in the United States. This dataset might include patients more susceptible to reaching a death state rather than staying in the cancerous state thus affecting the transition probabilities. The number of outliers and the possible effects of the dataset is something that needs to be looked into when considering the results from this paper, and could be an area of improvement in the future.

In the second simulation, as previously discussed it was attempted to use state transition based on picking a random number between 0 and 90. This was achieved using the following line of python code: simulState + min(random.randint(0, 90-simulState), random.randint(0, 90-simulState)). This code ensures as it takes the minimum of two random numbers that there is a lower probability of picking a higher number. However, this code does not provide a clear statistical distribution of numbers as it does pick the minimum of two random numbers.

Therefore, whilst it can be seen from the mean patient paths that this results in an aggressive state transition it is not known what kind of state transition this simulates. which would require future research and possibly more simulations to determine in which risk category the results of this random state transition would lie.

Because of the very fast increase towards severe states this simulation could be considered slightly biased towards the RL policy as a 10-yearly policy will always perform worse since there is a significant period between screenings allowing a patient to reach a cancerous or death state easily in the 10-yearly interval.

Furthermore, it can be seen from the box plot in Figure 5 that the second simulation has fewer amount of outliers for both the RL policy and the 10-yearly policy than the transition probability simulation. This could either be due to the higher variability seen for both policies or that the random state transition has less variety in its generated patient paths.

6.2 Future work

The RL-based policy could be a more cost-effective approach to colonoscopy screenings. However, to fully utilize the cost-saving potential of RL-based screening intervals the RL models and its outcomes should be properly implemented in a socio-economic environment. The simulation did indicate a better cost-effectiveness for the RL policy however, this was given a slightly lower mean QALY for the patient. Thus research on how to best implement the RL-generated policies in the healthcare environment should be addressed properly to utilize the cost-effectiveness of the RL model to its fullest without sacrificing the average QALY for the patients.

Furthermore, in the second simulation which simulated a more aggressive colorectal cancer, the RL policy proved to provide a much better outcome for patients however at a worse cost-effectiveness than the general policy. As for patients with a high-risk profile, this might be a more suitable option in terms of average QALY outcome it will have to be researched which risk-categories patient groups benefit most from using an RL-based policy.

As the transition dataset and the dataset on which the Q-learning was performed, is based on hospitals from the United States only a bias might exist. Therefore in the future, the study could be improved by simulating with a larger and more diverse dataset to see if the same results still hold true for a more diverse population.

Lastly, as the RL-based policy does provide potentially interesting use cases it will have to be thoroughly researched where the RL-based policy can best be socio-economically and logistically implemented to improve the cost-effectiveness for a general patient population and improve the overall screening effectiveness for high-risk patient groups.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my supervisor A. Asadi for the opportunity to work with his currently unpublished model and hopefully contribute to the research on colonoscopy screening policies as well as the guidance throughout the thesis.

DISCLOSURE ON THE USAGE OF AI

During the preparation of this work the author used Grammerly in order to check for spelling and grammar mistakes. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

REFERENCES

[1] Mahboubeh Madadi Amin Asad, Jaleh Soltanpour. Optimizing colorectal cancer screening tests with anytime point-based valu iteration algorithm unpublished manuscript. *tbd*, 2024.

[2] Florence Bénard, Alan N Barkun, Myriam Martel, and Daniel von Renteln. Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations. *World Journal of Gastroenterology*, 24(1):124–138, January 2018.

[3] National cancer institute. Screening tests to detect colorectal cancer and polyps, 2021.

[4] Jesse Clifton and Eric Laber. Q-learning: Theory and applications. *Annu. Rev. Stat. Appl.*, 7(1):279–301, March 2020.

[5] Fatih Safa Erenay, Oguzhan Alagoz, and Adnan Said. Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing amp; Service Operations Management*, 16(3):381–400, July 2014.

[6] Grace N Joseph, Farid Heidarnejad, and Eric A Sherer. Evaluating the cost-effective use of follow-up colonoscopy based on screening findings and age. *Comput. Math. Methods Med.*, 2019:2476565, February 2019.

[7] Iris Lansdorp-Vogelaar, Amy B Knudsen, and Hermann Brenner. Cost-effectiveness of colorectal cancer screening. *Epidemiol. Rev.*, 33(1):88–100, June 2011.

[8] Iris Lansdorp-Vogelaar, Marjolein van Ballegooijen, Ann G Zauber, J Dik F Habbema, and Ernst J Kuipers. Effect of rising chemotherapy costs on the cost savings of colorectal cancer screening. *J. Natl. Cancer Inst.*, 101(20):1412–1422, October 2009.

[9] Weiyu Li, Brian T. Denton, and Todd M. Morgan. Optimizing active surveillance for prostate cancer using partially observable markov decision processes. *European Journal of Operational Research*, 305(1):386–399, February 2023.

[10] Y Li, M Zhu, R Klein, and N Kong. Using a partially observable markov chain model to assess colonoscopy screening strategies – a cohort study. *Eur. J. Oper. Res.*, 238(1):313–326, October 2014.

[11] Mingyang Liu, Xiaotong Shen, and Wei Pan. Deep reinforcement learning for personalized treatment recommendation. *Stat. Med.*, 41(20):4034–4056, September 2022.

[12] Tomasz Skrzypczak, Anna Skrzypczak, and Małgorzata Skrzypczak. Implications of public interest in colonoscopy: Analysis of google trends data from 12 european countries. *Cureus*, 15(7):e42395, July 2023.

[13] Adam Yala, Peter G Mikhael, Constance Lehman, Gigin Lin, Fredrik Strand, Yung-Liang Wan, Kevin Hughes, Siddharth Satuluru, Thomas Kim, Imon Banerjee, Judy Gichoya, Hari Trivedi, and Regina Barzilay. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat. Med.*, 28(1):136–143, January 2022.

A APPENDIX

A.1 Python code Q-learning

```
##Bayesian q-learning
import numpy as np

class BayesianQLearning:
    def __init__(self, num_states, num_actions,
         initial_alpha=1.0, min_alpha=0.005,
         decay_rate_alpha=0.01,
                 initial_beta=0.01, min_beta = 0.001,
                      decay_rate_beta = 0.01):
        self.num_states = num_states
        self.num_actions = num_actions
        self.initial_alpha = initial_alpha
        self.min_alpha = min_alpha
        self.decay_rate_alpha = decay_rate_alpha
        self.initial_beta = initial_beta
        self.min_beta = min_beta
        self.decay_rate_beta = decay_rate_beta
        self.Q_mean = np.zeros((num_states, num_actions))
              # Mean of the posterior
        self.Q_var = np.ones((num_states, num_actions))
             # Variance of the posterior (initialized to
             1)
        self.N = np.zeros((num_states, num_actions))  #
             Count of updates (for variance calculation)

    def update(self, state, action, reward, next_state,
         episode):
        # Update count
        self.N[state, action] += 1

         # Calculate the current learning rate
        alpha = self.alpha_rate(episode)
        beta = self.beta_rate(episode)

        # Bayesian update of Q-value mean
        old_mean = self.Q_mean[state, action]
        old_var = self.Q_var[state, action]
        new_mean = old_mean + alpha * (reward - old_mean)
        new_var = old_var * (1 - beta)

        self.Q_mean[state, action] = new_mean
        self.Q_var[state, action] = new_var

    def select_action(self, state):
        # Sample from the posterior distribution
        sampled_values = np.random.normal(self.Q_mean[
             state], np.sqrt(self.Q_var[state]))
        return np.argmax(sampled_values)
    def bayes_update(self,state,action,timeHorizon):
        return updateState(state,action,timeHorizon)
    #calculate decreasing rate
    def alpha_rate(self, episode):
        return max(self.min_alpha, self.initial_alpha *
             np.exp(-self.decay_rate_alpha * episode))
    #calculate decreasing rate
    def beta_rate(self,episode):
        return max(self.min_beta, self.initial_beta * np.
             exp(-self.decay_rate_beta * episode))



# Initialize Bayesian Q-learning agent
Qenv = BayesianQLearning(n_states, n_action)

# Training loop
n_episodes = 20000
for episode in range(n_episodes):
    state = random.choice(Polyp_states)  # Set initial
         state for each episode
    for t in range(T):
        if state >= Death_state:
            break

        # Select action based on posterior sampling
        action = Qenv.select_action(state)

        # Perform action and observe reward and next
             state
        reward = rewards[state,action,t]
        next_state = Qenv.bayes_update(state, action, t)
              # Update state
        # Update Q-values (Bayesian update)
        Qenv.update(state, action, reward, next_state,
             episode)

        # Update current state to next state
        state = next_state

print("Q-table learned (posterior mean)")
print(Qenv.Q_mean)
#safe for later use
Qlearned = Qenv.Q_mean
```

A.2 Python code simulation

```
#simulation
def simulatePatient(policy, initState):
    path  = np.full(25, np.nan) #define list of 25 items
    path[0] = initState
    reward = 0 #intialize reward
    #if isinstance(initState, list):
     #   path += initState
    #else:
     #   path.append(initState)
    #define currState
    currState = initState
    colonoscopyAmount = 0
    #simulate for time epochs
    lastObs = 0
    t = 0
    while t < T:
        #first determine action for policy
        simulaction = determineAction(currState,t,policy)
        #if colonoscpy is performed increase colonoscpy
             count and register observation
        if(simulaction == 1):
            colonoscopyAmount +=1
        #update state
        path[t] = currState
        currState = updateSimulationState(currState,
             simulaction,t)

        #add to path
        reward += rewards[currState,simulaction,t]
        t += 1

    return reward,colonoscopyAmount,path,currState
```