＜ 修 士 論 文 ＞

前立腺がんの腫瘍増殖率を最小化する動的治療計画のための深層強化学習

滋 賀 大 学 大 学 院
デ ー タ サ イ エ ン ス 研 究 科
デ ー タ サ イ エ ン ス 専 攻

修了年度：２０２２年度
学籍番号：６０２１１１６
氏　　名：朱　澤胤
指導教員：岩山　幸治
提出年月日：２０２３年１月１０日

目次

1 はじめに 1

1.1 前立腺がん 1
1.2 前立腺がんの治療法 1
1.3 本研究の目的 2

2 先行研究 3

2.1 数理モデル 3
2.2 データ 4

3 動的治療計画 6

4 深層強化学習 7

4.1 強化学習 7
4.2 Deep Q Learning(DQN) 8

5 深層強化学習による前立腺がん治療のシミュレーション 11

5.1 シミュレーションデータの作成 11
5.1.1 パラメータの算出 11
5.1.2 パラメータのオーバーサンプリング 12
5.2 本研究における深層強化学習の設定 13
5.3 モデルの適用 14

6 分析結果 15

6.1 オフライン Q 学習の結果 15
6.2 オンライン Q 学習の結果 20

7 考察 26

8 結論 27

謝辞 28

引用・参考文献 28

1 はじめに

1.1 前立腺がん

前立腺がんの 2020 年の罹患数は全世界で約 141 万人であり，全種類のがんの中で 7.3%を占め第 4 位であった．前立腺がんの罹患率は特に先進国で高く，発展途上国の約 5 倍となっている. 前立腺がんによる死亡数は年間約 38 万人で全種類のがんの 3.8%を占め第 8 位であった[1]．また，日本においては男性のがん罹患予測(2016 年)で第 1 位となるなど，近年前立腺がんの罹患率が急上昇している[2]．日本において前立腺がんの罹患率が増えている理由は 3 点ある．1 点目は日本人の高齢化である．前立腺がんは主に 60 歳以上に多く見られ，とくに 80 歳以上では半数以上に潜在性の前立腺がんがあると言われている. そのため，日本では高齢化に伴って前立腺がんの患者が増えてきている. 2 点目は食生活の欧米化により動物性脂肪を多く摂るようになったことが前立腺がん発症に影響を及ぼしていると考えられている．3 点目は PSA 検査の普及により，直腸内触診や超音波検査では発見することが困難であった，症状が現れない早期のがんを発見することが可能になったことである[2].

PSA（Prostate-Specific Antigen）とは前立腺から分泌されるタンパク質分解酵素の一種であり，前立腺がんに罹患すると血液中の量が増加することから，前立腺がんの腫瘍マーカーとして使われている．PSA の正常値は 4.0ng/mL 以下であり，4.0ng/mL を超えると前立腺がんに罹患している可能性が高いとされる．前立腺がん細胞の量を直接計測することは困難であるため，代わりに PSA が前立腺がん細胞の量を表す指標となっている. 例えば前立腺がんの増殖モデルに関する先行研究においては, PSA に基づいてモデルを構築している[3].

1.2 前立腺がんの治療法

前立腺がんの治療法には，手術治療，放射線治療，ホルモン療法がある．前立腺がんが早期であり，他の臓器などに転移がなければ手術治療，放射線治療のいずれかを選択することが可能である. しかし進行が見受けられ，他の臓器への転移などが生じている場合や，年齢，身体的な問題で手術など侵襲の大きい治療が受けられない場合にはホルモン療法が行われる.

前立腺がんは，男性ホルモンであるアンドロゲンが前立腺のアンドロゲン受容体と結合することによって増殖することが知られている[2]．ホルモン療法はアンドロゲンの分泌や働きを妨げる薬によって前立腺がんの増殖を抑える治療である[4]．また，ホルモン療法には継続的にアンドロゲンを抑制するもの(継続的な投薬)と間欠的にアンドロゲンを抑制するもの(間欠的な投薬)の 2 種類がある．継続的な投薬はアンドロゲンに依存せずに増殖するがん細胞の増殖を早めてしまい，結果的にがん細胞の増殖が

再燃すると考えられる. 一方で, 間欠的な投薬の場合は再燃問題を解決する可能性があり, 継続的な投薬よりも治療効果が高いという研究結果がある[5].

間欠的な投薬は継続的な投薬に比べ腫瘍増殖率を抑えることに加え, QOL の向上や医療費の削減に貢献すると期待される. しかしながら, 投薬開始および投薬中止のタイミングを個々の患者に対して決めるためのプロトコルは未だに確立されていない[6].

1.3 本研究の目的

本研究では前立腺がん細胞の増殖モデルを仮定せず, 深層強化学習によって得られたモデルに従って投薬の意思決定を行うことを提案する. シミュレーションにより, 深層強化学習に基づく投薬の意思決定が先行研究の治療ガイドラインに従った意思決定に比べがんの増殖を抑えられることを示す. また, 提案手法は増殖モデルを仮定しないため, 前立腺がんに限らず再燃をするタイプの課題に対しても同じように深層強化学習を用いて意思決定を支援することが有効であると期待される.

2 先行研究

先行研究では前立腺がん細胞の増殖を正確に予測できる数理モデルが作成されている[7]．本研究ではこのモデルにしたがって任意の投薬スケジュールのもとでの前立腺がん細胞の増殖のシミュレーションを行う．

2.1 数理モデル

先行研究では前立腺がん細胞について①投薬によって抑制可能な前立腺がん細胞，②投薬によって抑制できないが，投薬を中止すると抑制可能ながん細胞に変化する前立腺がん細胞，③投薬によって抑制できず，投薬を中止しても変化することがない前立腺がん細胞という 3 種類の細胞の存在を仮定している(図 1)．上記①の前立腺がん細胞はアンドロゲン依存がん細胞，②と③の前立腺がん細胞はアンドロゲン非依存がん細胞と呼ばれる[7]．

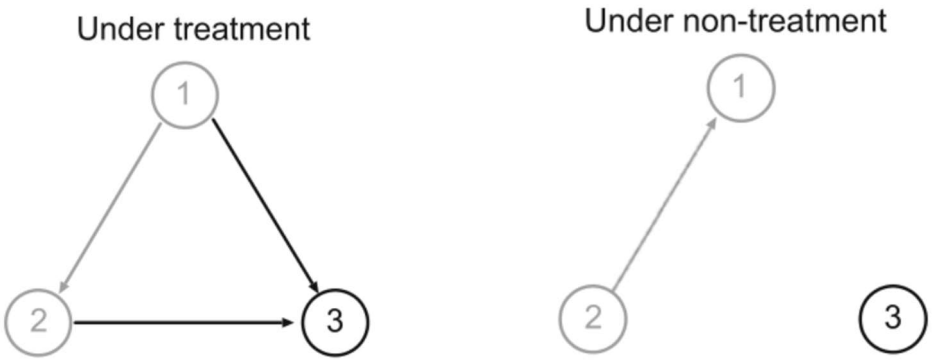

図 1．3 種類のがん細胞の増殖のダイアグラム．左が投薬期間，右が投薬中止期間に対応する．投薬期間中はがん細胞①は②と③に変化し，②は③に変化していく．投薬中止期間中はがん細胞②は①に変化する．

時刻tにおける①，②，③のがん細胞の量をそれぞれ$x_1(t)$，$x_2(t)$,$x_3(t)$とする．このとき，投薬期間と投薬中止期間の前立腺がん細胞の量はそれぞれ式 1, 2 に示す差分方程式に従って変化する．

$$\begin{pmatrix} x_1(t+\Delta t) \\ x_2(t+\Delta t) \\ x_3(t+\Delta t) \end{pmatrix} = \begin{pmatrix} d^1_{1,1} & 0 & 0 \\ d^1_{2,1} & d^1_{2,2} & 0 \\ d^1_{3,1} & d^1_{3,2} & d^1_{3,3} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} \tag{1}$$

$$\begin{pmatrix} x_1(t+\Delta t) \\ x_2(t+\Delta t) \\ x_3(t+\Delta t) \end{pmatrix} = \begin{pmatrix} d^0_{1,1} & d^0_{1,2} & 0 \\ 0 & d^0_{2,2} & 0 \\ 0 & 0 & d^0_{3,3} \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} \tag{2}$$

ここで，パラメータの集合$\{d_{i,j}^m\}$はそれぞれの種類のがん細胞の増殖率及び異なる種類の細胞への変化率を表す．患者から測定された PSA の系列とモデルにおけるがん細胞の総量$x_1(t)+x_2(t)+x_3(t)$との間の誤差を最小化することで，初期値$x_1(0), x_2(0), x_3(0)$及びパラメータの集合$\{d_{i,j}^m\}$の値を求め，任意の投薬スケジュールのもとでのがん細胞の量を予測することが可能になる[7].

2.2 データ

先行研究の間欠的な投薬において投薬の開始と中止の決定に用いられたガイドライン[7][8]を図 2 に示す．36 週間継続で投薬したら投薬を中止し，PSA が 10 を超えた時点で投薬を再開する．PSA の値は 28 日ごとに測定し，投薬期間に PSA が 3 回連続で増加した場合，あるいは投薬開始後 24 週目と 32 週目の PSA が 4 未満の場合，試験を中断する．患者 72 名について，このガイドラインに従って治療を行い，PSA の測定が行われた．

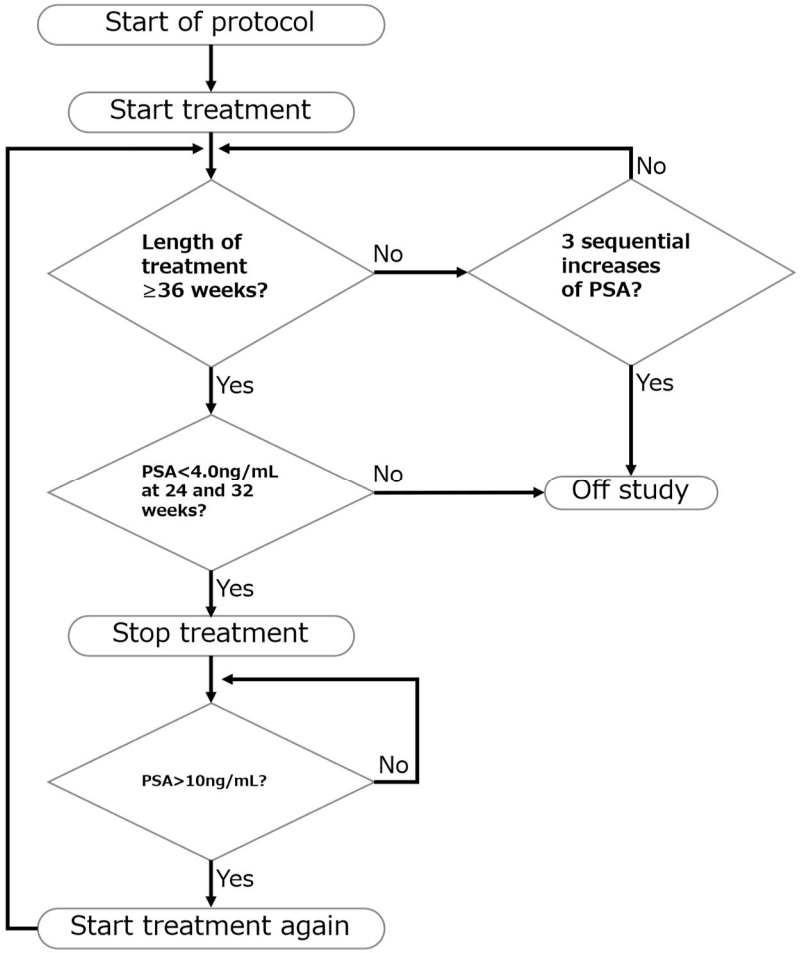

図 2. 先行研究[7][8]で用いられた間欠的な投薬の治療ガイドライン

ガイドラインに従って治療を行った患者のうち一名についてモデル (式 1, 2) を当てはめた結果を図 3 に示す．間欠的な投薬のもとでの PSA の変化はモデルによる予測と概ね一致していることがわかる．その一方で，灰色の実線が示すように，継続的に投薬した場合, PSA 値はいったん減少するが，ある時点を境に薬を与えているのにもかかわらず PSA 値が増加している様子が分かる．

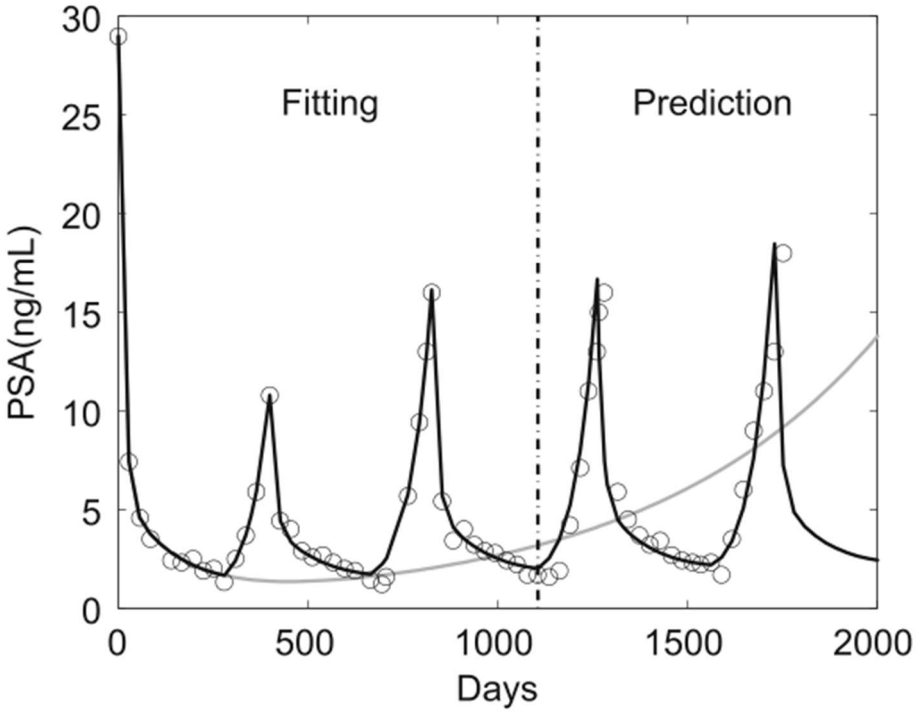

図 3. PSA の実測とモデルによる予測．PSA の測定値を丸で，この患者に行った投薬スケジュールのもとでのモデルの推測を黒い実線で表す．図中に黒い破線で示した時点までのデータでモデルのパラメータの最適化を行っており，破線より右側の区間はモデルによる予測に対応する. また，継続的な投薬を行ったと仮定した場合のモデルの予測値を灰色の実線で表す[7].

3　動的治療計画

個別化医療は個々の患者の情報を体系的に利用し，その患者の健康管理を最適化することを目的としたパラダイムである．このような医療パラダイムの動機として，特定の治療に対する患者の反応が，主要な結果と副作用の両方の観点から異なることが多いという事実がある．このような患者の反応の不均一性を考慮して，多くの医療研究者は，画一な医療から個別化医療という近代的かつ論理的なアプローチに移行した[9]．個別化医療のメリットは，治療に対するコンプライアンスやアドヒアランス[1]の向上，最適な治療法を選択することによる患者ケアの向上，医療費全体の削減などが挙げられる[9]．患者を効果的かつ長期的にケアするためには，慢性期医療モデル(Chronic care model)に従って，継続的な医療介入が必要である．慢性期医療モデルは患者の反応，アドヒアランス，負担，副作用，嗜好などの継続的な測定結果に応じて，治療の種類，量，タイミングを個別化して治療を行い，患者のニーズに応じたパーソナライゼーションを重視する．慢性医療モデルを実践する臨床家は，一度にすべての治療法を決定する(静的治療)のではなく，個々の患者のケースヒストリーに基づいて，患者の転帰を最適化するために次に何をすべきかを逐次決定していく(動的治療)．一連の治療を検討する主な動機は，治療に対する反応の患者間のばらつきが大きいこと，再発の可能性があること，並存疾患の存在または出現，時間的に変化する副作用の重症度，集中的な治療が不要な場合のコストと負担の軽減などが挙げられる．従来の臨床家向けの診療ガイドラインは主に「専門家の意見」に基づいているが，慢性期医療モデルではこれらの体制をより客観的で証拠に基づいたものにすることを提唱している[9]．Parmigiani(2002)が主張するように，意思決定理論的な考え方の医学への主な貢献は，意思決定に関連する定量的な情報を集め，整理し，統合するプロセスに構造を提供し，目標を正式に定義することにある[10]．このことは，一般の人々への伝達が困難であるにもかかわらず，医学的意思決定における意思決定理論的形式論の役割を正当化するものである[9]．

動的治療計画はあるステージまでのある患者個人の特徴量と治療歴をインプットとして，そのステージにおいて推薦される治療(治療タイプ，投薬量，投薬タイミング)を出力とする．形式的に表現すると，ステージ$j(1 \leqq j \leqq K)$において，エージェントは状態 O_j を観測し，行動 A_j をとる．行動の結果，報酬 Y_j を得る．次のステージに移り，O_{j+1} が観測される．$H_j = (O_1, A_1, \ldots, A_{j-1}, O_j)$としたときに，動的治療計画とは$H_j \to A_j$ を決定する方策d_jであり，$d_j = (A_j | H_j)$と表記する[9]．以上より，前立腺がんの間欠的な投薬における課題の解決は個別化医療を実現させるための慢性期医療モデルを利用でき，慢性期医療モデルを構築するための動的治療計画のフレームワークは強化学習と非常に似ているため，本研究では深層強化学習を用いて課題解決を試みた．

---

[1] 患者が治療方針の決定に賛同し積極的に治療を受けること．

4　深層強化学習

4.1 強化学習

強化学習の基本的な構図は図 4 のようにまとめることができる.

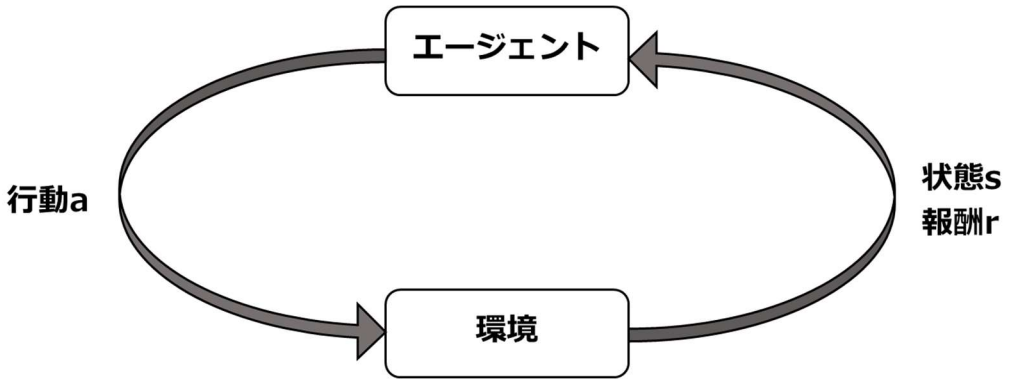

図 4. 強化学習の構図

強化学習は機械学習の中で，特に意思決定を学習する方法である．意思決定の主体であるエージェントが，ある環境において目的を達成するためにどのような意思決定をして行動するべきかという方策を学習する．また，ある状態で行動をとることで得られる即時報酬ではなく，行動を繰り返した後に得られる長期的な収益の最大化を目指す.

多くの強化学習の問題では,環境をマルコフ決定過程として定式化する.マルコフ決定過程では，現在の状態s_tと行動a_tのみに依存して，次の時刻$t+1$の状態s_{t+1}が決まる．マルコフ決定過程は以下の構成要素によって定義される:

- 状態集合: 環境内でエージェントが観測する状態の集合$S = \{s_1, s_2, s_3, \dots, s_N\}$
- 行動集合: エージェントが環境に対して起こす行動の集合$A = \{a_1, a_2, a_3, \dots, a_M\}$
- 状態遷移確率: 状態s_tで行動a_tを取って次の状態s_{t+1}に遷移する確率$P(s_{t+1}|s_t, a_t)$
- 報酬関数:状態s_tで行動a_tを取って状態s_{t+1}を観測して得る報酬$r_{t+1} = r(s_t, a_t, s_{t+1})$

エージェントがある状態sで行動aを選択する確率$\pi(a|s)$を方策という．強化学習の目的は一連の状態観測と行動の繰り返しによって獲得する報酬の総和である収益を最大化する方策を見つけることである．ただし，単純な報酬和をとると，無限ステップの行動で発散してしまうため, 0~1 の割引率γを導入して，以下の割引収益を考える．

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} \dots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

割引収益を最大化することが強化学習の目的であるが，未来における報酬を知ることはできないため，割引収益の期待値を取った価値関数を導入する．状態sで方策πに従って行動aをとった時の価値関数を行動価値関数と呼び，$Q^{\pi} = E[R_{t+1}|S_t = s, A_t = a]$と表記する．

行動価値関数Q^{π}は以下のように分解でき，これをベルマン方程式と呼ぶ．

$$Q^{\pi}(s,a) = \sum_{s' \in S} P(s'|s,a)\left(r(s,a,s') + \gamma \sum_{a' \in A} \pi(a'|s')\, Q^{\pi}(s',a')\right)$$

上式より，行動価値関数Q^{π}は状態sで行動aをとった後に得られた報酬と遷移先の状態の価値関数の和の期待値で表される．

強化学習の手法の一つにベルマン方程式に基づいて行動価値関数の学習を行う Q 学習がある．以下 Q 学習について説明する．

TD ターゲット（Temporal Difference Target）と呼ばれる量を以下で定義する．

$$Y_t = r_{t+1} + \gamma \max_{a_{t+1}} \hat{Q}^{\pi}(s_{t+1}, a_{t+1})$$

ここで$\hat{Q}^{\pi}$は行動価値関数の推定である．ベルマン方程式より，状態s_tで行動a_tをとった後に得られた報酬r_{t+1}(実測値)と状態s_{t+1}以降に得られる最大割引収益(予測値)の和の期待値が行動価値関数であるため，TD ターゲットは得られた報酬と行動価値関数の推定に基づくその近似であり，教師信号の役割を果たす．

行動価値関数の推定は以下のように更新される．

$$\delta_t = Y_t - \hat{Q}^{\pi}(s_t, a_t)$$
$$\hat{Q}^{\pi}(s_t, a_t) \leftarrow \hat{Q}^{\pi}(s_t, a_t) + \alpha\delta_t$$

δ_tは TD 誤差(Temporal Difference Error)と呼ばれるもので，TD ターゲットY_tが行動価値関数の推定$\hat{Q}^{\pi}(s_t, a_t)$に比べてどれだけずれているかを表す．αは割引率を表す．

4.2 Deep Q Learning (DQN)

DQN は深層強化学習の一種で，行動価値関数$\hat{Q}^{\pi}(s,a)$をニューラルネットワークで直接推定する手法である．このニューラルネットワークを Q ネットワークと呼ぶ．Q ネットワークのイメージは図 5 のとおりである．（行動の種類がa_1, a_2の 2 つの場合）

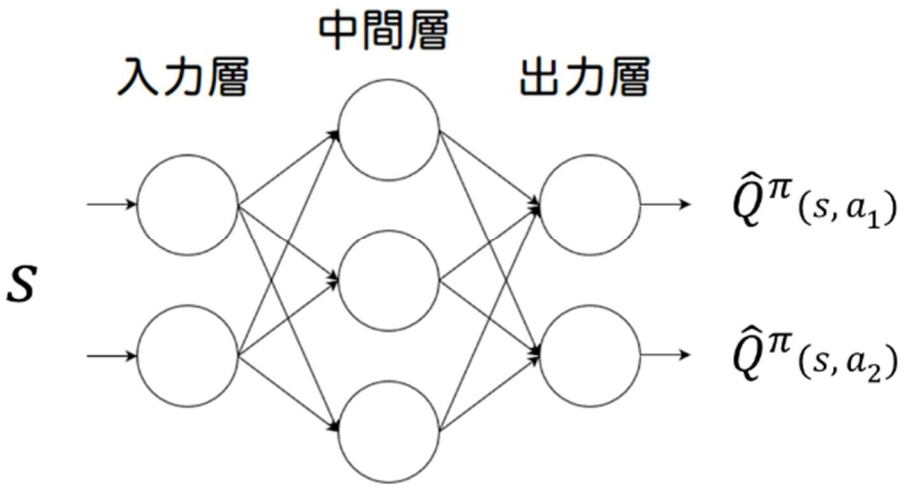

図 5．DQN のイメージ

状態Sを入力として，出力は各行動を取った場合の行動価値である．

Q ネットワークの学習では損失関数を最小化するように重みを更新する．

本研究では，損失関数について Huber Loss を用いた．Huber Loss の定義を式 3 に示す．

$$L_{\delta}\big(y,f(x)\big)=\begin{cases}\frac{1}{2}\big(y-f(x)\big)^2, \text{for } |y-f(x)|\le\delta,\\ \delta(|y-f(x)|-\frac{1}{2}\delta), \text{ otherwise.}\end{cases} \tag{3}$$

ここで，Lは実測値yと予測値$f(x)$の損失関数を表し，δは損失関数が 2 次関数から線形に変わる点を表す．本研究では$\delta=1$とした．DQN の学習においては，目的変数yは TD ターゲットY_t，関数$f(x)$は Q ネットワークの出力する行動価値の推定値に対応する．

最適化手法は Adam を用いた．Adam は勾配降下法における振動を抑制するために考案された最適化手法である[11]．Adam による重みwの更新式を式 4, 5, 6 に示す．

$$m_t=\beta_1\cdot m_{t-1}+(1-\beta_1)\cdot g_t \tag{4}$$

$$v_t=\beta_2\cdot v_{t-1}+(1-\beta_2)\cdot g_t^2 \tag{5}$$

$$w_t=w_{t-1}-\frac{\alpha\cdot m_t}{\sqrt{v_t+\epsilon}} \tag{6}$$

ここで，$\beta_1,\beta_2,\alpha,\epsilon$はそれぞれパラメータ，$w_t$は$t$回目の更新における Q ネットワークの重みのパラメータである．また，g_tは損失関数の勾配を表す．式 4, 5, 6 より，勾配の移動平均m_tを求めることで振動を抑えていること，勾配の 2 乗の移動平均v_tで更新量を割ることで学習率を調整していることがわかる．本研究では$\beta_1=0.9,\beta_2=0.999,\alpha=0.001,\epsilon=10^{-7}$とした．

DQN のミニバッチ学習では，エージェントの行動によって得られるデータは強い系列相関を持つため，これを使って逐次的に学習を行うと,性能が低下する恐れがある．そこでデータの偏りを防ぐために瞬時的な経験$(s_t, a_t, r_{t+1}, s_{t+1})$をリプレイメモリーという集合に保存する．学習時にはリプレイメモリーからランダムに選ばれた経験をミニバッチとして用いる．これにより学習データ系列内の相関が消え，学習が安定化すると期待される．

勾配降下法でパラメータを更新する際に，標的信号となる TD ターゲットにおける$\hat{Q}^{\pi}(s_{t+1}, a_{t+1})$は学習に伴い値が変化していくため，学習が安定しなくなる．そこで，学習に伴うターゲット自体の変化を抑えるために Q ネットワークとは別にターゲットネットワークを構築する．行動価値$\hat{Q}^{\pi}(s_t, a_t)$を計算する際は Q ネットワークを, TD ターゲットを計算する際はターゲットネットワークを用いる．

Q ネットワークの重みをw，ターゲットネットワークの重みをw^{-}とすると，ターゲットネットワークの重みは$w^{-} \leftarrow \tau w + (1-\tau)w^{-}$と更新する．$\tau$は二つのネットワークの重みを近づける割合で，小さく設定することでターゲットネットワークの更新がゆっくりとなり，学習が安定する．本研究では$\tau = 0.001$とした．

Q ネットワークを構築するための学習方法はオフライン Q 学習とオンライン Q 学習に分けられる．本研究ではその両方についてそれぞれ Q ネットワークを構築した．オフライン Q 学習は過去に収集したデータのみを用いて行うものであり，学習に実環境とのインタラクションが求められないため，医療分野において安全面のリスクがオンライン Q 学習に比べて少ない．ただし，データセット分布外の行動を過大評価してしまう(分布シフト)オフライン強化学習特有の問題が潜在し，意思決定がうまくいかない可能性がある．その一方で，オンライン Q 学習は実環境とのインタラクションが求められ，実環境に作用しながら学習を行うため，安全面のリスクが比較的に大きいが, Q ネットワークに従って選択した行動に対して観測されたデータを学習に用いることができる．

医療分野では一般的な強化学習と違い，学習データの取得時にも安全性が求められるため，本研究ではオフライン Q 学習に用いるデータを，先行研究のガイドラインに従った間欠的な投薬治療によって取得する状況を想定している．こうして得られる学習データは多様性に乏しく，オフライン Q 学習で得られるモデルの性能も低下することが懸念される．そこで，本研究では実用性を考慮し，オフライン Q 学習で一旦学習を行い，学習後の Q ネットワークを実環境に作用させ，新たにデータが収集されるたびにリプレイメモリーに保存し Q ネットワークの更新を行った．オフライン Q 学習をあらかじめ行うことで安全面のリスクを抑えつつ，新たなデータでモデルを更新することでより性能の高い Q ネットワークを構築できると期待される．

5　深層強化学習による前立腺がん治療のシミュレーション

5.1 シミュレーションデータの作成

5.1.1　パラメータの算出

本研究ではリアルワールドの患者に対してではなく，擬似的な患者に対して意思決定のシミュレーションを行う．72 人分のデータが記載されている現実のデータセットを入手した[7]．このデータセットは各患者について 28 日ごとに測定された PSA 検査値，投薬有無の情報が含まれている．先行研究の数理モデルに従って，この二つの情報から各患者のがん増殖パラメータ(PSA 初期値とがん細胞の増殖パラメータ)を推測した．一部の患者について求めたパラメータと実際の投薬スケジュールの下でのモデルの推測結果と PSA の実測値を図 6, 7 に示す．患者 72 名中 69 名については，図 6 に示したようにモデルによる予測と実測は同様のふるまいを見せたが，図 7 に示したように 3 名についてはモデルによる予測は実測と全く異なるものとなった．

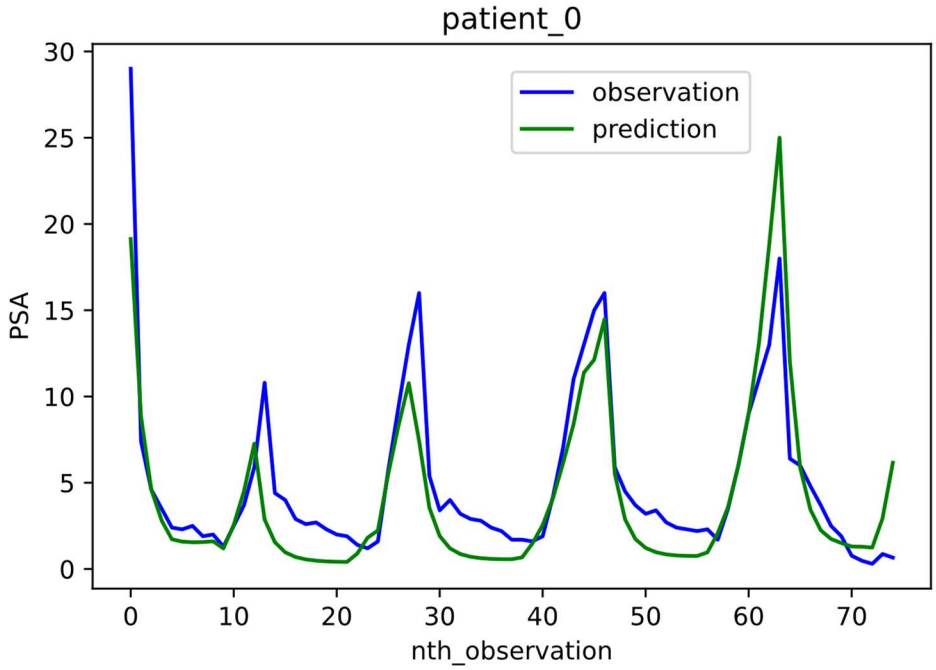

図 6．一部の患者についての PSA の推移．横軸が PSA 値の観測回数(28 日ごとに 1 回観測)，縦軸が PSA 観測値を表す．observation は実測値, prediction はモデルによる予測を示す．また, PSA の予測が当てはまった例(72 人中 69 人)．

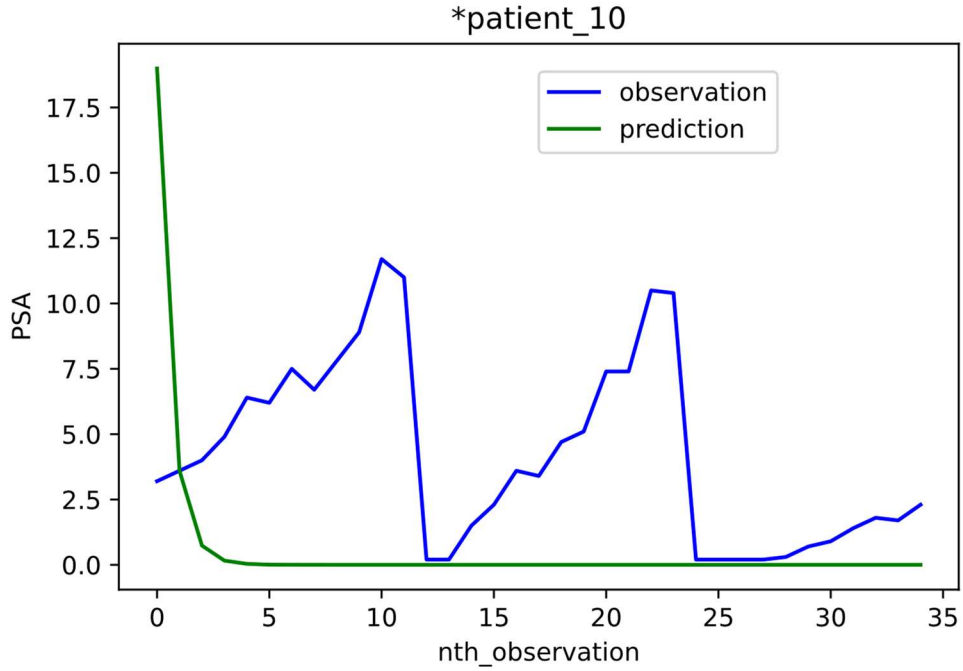

図 7. 一部の患者についての PSA の推移. 横軸が PSA 値の観測回数(28 日ごとに 1 回観測), 縦軸が PSA 観測値を表す. observation は実測値, prediction はモデルによる予測を示す. また, PSA の予測が当てはまらなかった例(72 人中 3 人)である.

5.1.2 パラメータのオーバーサンプリング

より多様な患者についてシミュレーションを行うため, PSA の実測にモデルがよく当てはまった 69 人について, SMOTE[12]によりオーバーサンプリングを行った. オーバーサンプリングの結果得られた 1035 人分の疑似患者のパラメータのもとで, ガイドラインに従って治療を行った場合の PSA の推移をシミュレーションし, 3 年分の PSA の観測履歴($\mathrm{PSA}_1, \mathrm{PSA}_2, \ldots$)と投薬有無の履歴($\mathrm{Dose}_1, \mathrm{Dose}_2, \ldots$)を得た. これらの 1035 人分のデータのうち 835 人分を学習データ, 100 人分をバリデーションデータ, 100 人分をテストデータとした.

5.2 本研究における深層強化学習の設定

状態，行動，報酬，割引率についての設定は以下の通りである．

$$s_t = (\mathrm{PSA}_{t-n+1}, \ldots, \mathrm{PSA}_t, \mathrm{Dose}_{t-n+1}, \ldots, \mathrm{Dose}_{t-1})$$
$$a_t = \mathrm{Dose}_t$$
$$r_{t+1} = \exp(-\mathrm{PSA}_{t+1})$$
$$\gamma = 0.99$$

状態には始点tから数えて合計nステップ分の PSA 観測履歴と$n-1$ステップ分の投薬履歴が含まれている．行動はtステップ目の投薬状況としている．例えば$n=5$の場合，5 回分の PSA 観測履歴と 4 回分の投薬履歴をもとに次に選ぶべき行動を決定しようというものである．報酬は次に観測される PSA 値が小さいほど大きくなるように設定することで，腫瘍増殖率の最小化を目指す．

Q ネットワークの入力層のノード数は PSA 履歴n回分と投薬履歴$n-1$回分で合計$2n-1$，中間層は 3 層で各層のノード数は 32 とした．出力層のノード数は投薬有り無しの行動に応じて 2 つと設定した．

バリデーションデータの 100 人分の擬似患者を用いてハイパーパラメータである PSA 履歴のステップ数nの調整を行った．表 1, 2 に$n=5,6,7$ それぞれについて深層強化学習によって投薬スケジュールを決めた場合の PSA の最大値の 25, 50, 75, 90%点を示す．オフラインおよびオンライン Q 学習において，75%点の値が共に低く，Q ネットワークへ入力する観測数が少ないため治療開始後の最も早い段階で深層強化学習による投薬の意思決定を利用できることから，以降は$n=5$とする．

表 1 $n=5,6,7$それぞれの場合における PSA 最大値の 25%, 50%, 75%, 90%点
（オフライン Q 学習）

	25%	50%	75%	90%
$n=5$	1.38	2.07	3.69	15.47
$n=6$	5.58	8.97	10.03	10.69
$n=7$	1.65	2.51	11.27	14.37

表 2 $n=5,6,7$それぞれの場合における PSA 最大値の 25%, 50%, 75%, 90%点
（オンライン Q 学習）

	25%	50%	75%	90%
$n=5$	0.18	0.36	0.80	1.41
$n=6$	0.18	0.43	1.26	11.18
$n=7$	0.18	0.32	0.67	1.39

5.3 モデルの適用

テストデータである 100 人分の擬似患者に対して学習した Q ネットワークを適用する．PSA の値は先行研究と同様に 28 日ごとに観測し，次の観測までの期間，投薬を続けるか中断するかの意思決定を行うものとする．Q ネットワークの入力に PSA 観測履歴 5 回分と投薬履歴 4 回分が必要なため，各患者について最初に継続的に投薬を行ったもとで 5 回 PSA を観測する．こうして得られた 5 回分の PSA 観測履歴と 4 回分の投薬履歴を Q ネットワークに入力し，投薬した場合と投薬しなかった場合の行動価値行動価値を出力する．Q ネットワークの出力した行動価値が大きい方の行動を取り続けるように意思決定を行う(Greedy な方策）という方策に従って 10 年間の投薬シミュレーションを行った．また，意思決定の段階では想定外の出力によって PSA が急激に増加しないようにガイドラインの方策の一部を導入した．具体的には PSA が 10 を超えると強制的に投薬に切り替えるようにすることによって保守的な意思決定を行う．

オンライン学習における Q ネットワークの更新は，テストデータの中のある擬似患者に対して最新の Q ネットワークを適用し続け(この間は更新を行わない)，得られた一連の観測データをリプレイメモリーに追加し Q ネットワークを更新する．これをテストデータの人数分行う．

6　分析結果

6.1 オフライン Q 学習の結果

テストデータ 100 人の擬似患者に対して第 5 章で学習した Q ネットワークを適用した場合と，ガイドラインに従って間欠的な投薬を行った場合のそれぞれについて，10 年分の PSA の推移結果を得た．ガイドラインで試験期間満了(10 年経過する)前に間欠投薬を中止と判断された患者とそうでない患者の例をそれぞれ図 8, 9 と図 10, 11 に示す．横軸が PSA 値の観測回数(28 日ごとに 1 回観測)，縦軸が PSA 観測値を表す．図 8, 9 はガイドラインに従って間欠投薬中止と判断され，その後投薬を継続した患者の例で，図 8 の例から提案手法(proposed)と比べてガイドライン(guide)の方は最終的に PSA 値が再燃してしまっていることが分かる．図 9 の例は提案手法もガイドラインも最終的に PSA 値が再燃してしまっているが，提案手法では PSA 値の増加がガイドラインに比べて遅いことが分かる．図 10, 11 はガイドラインに従って間欠投薬が最後まで続いた患者例で，図 10 の例は提案手法の方が PSA を低く抑えられていることが分かる．図 11 の例は提案手法もガイドラインもほぼ同じ程度に PSA を低く抑えられていることが分かる．

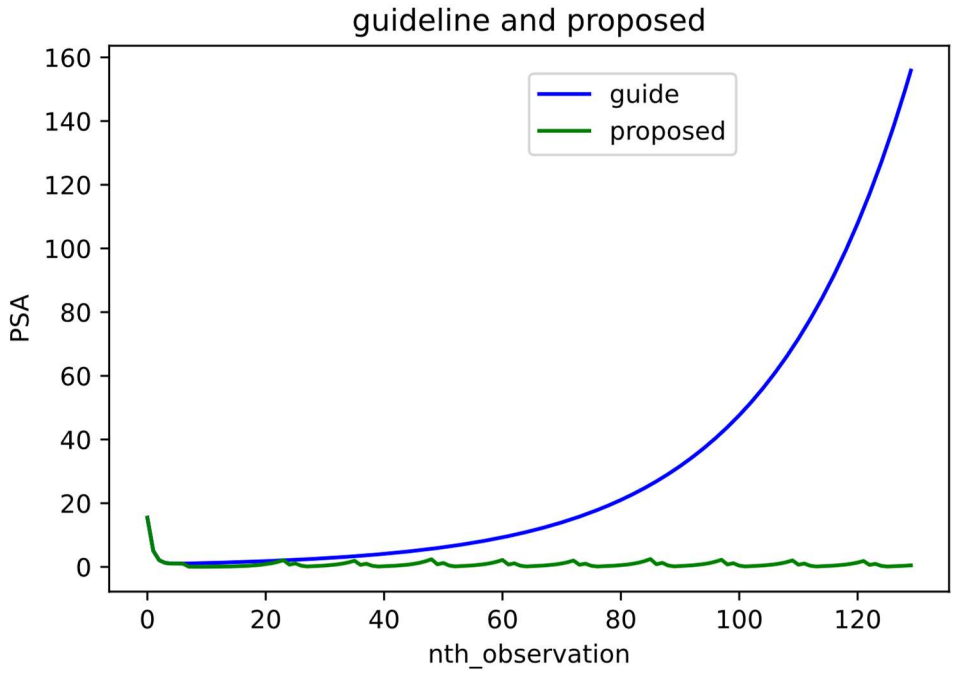

図 8.　ガイドラインで間欠投薬が中止になった後投薬を継続した患者の例 1

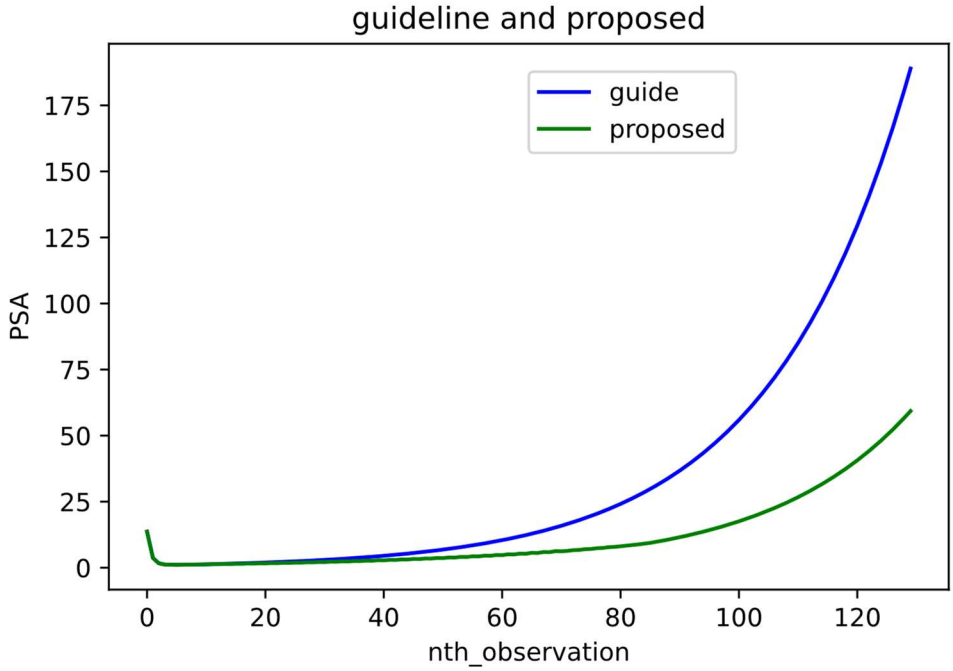

図 9. ガイドラインで間欠投薬が中止になった後投薬を継続した患者の例 2

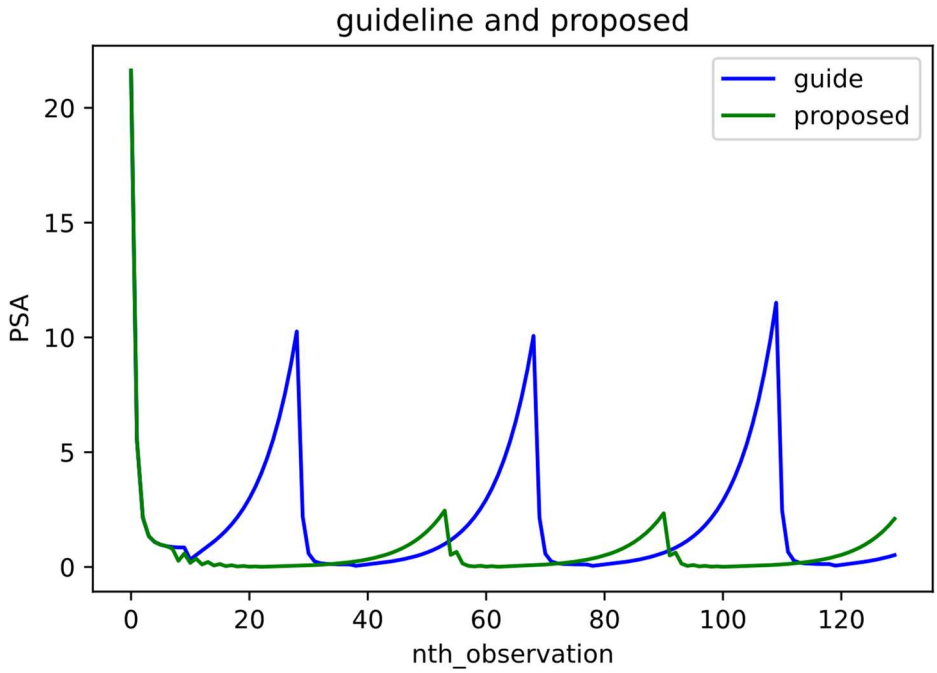

図 10. ガイドラインで間欠投薬が中止にならなかった患者の例 1

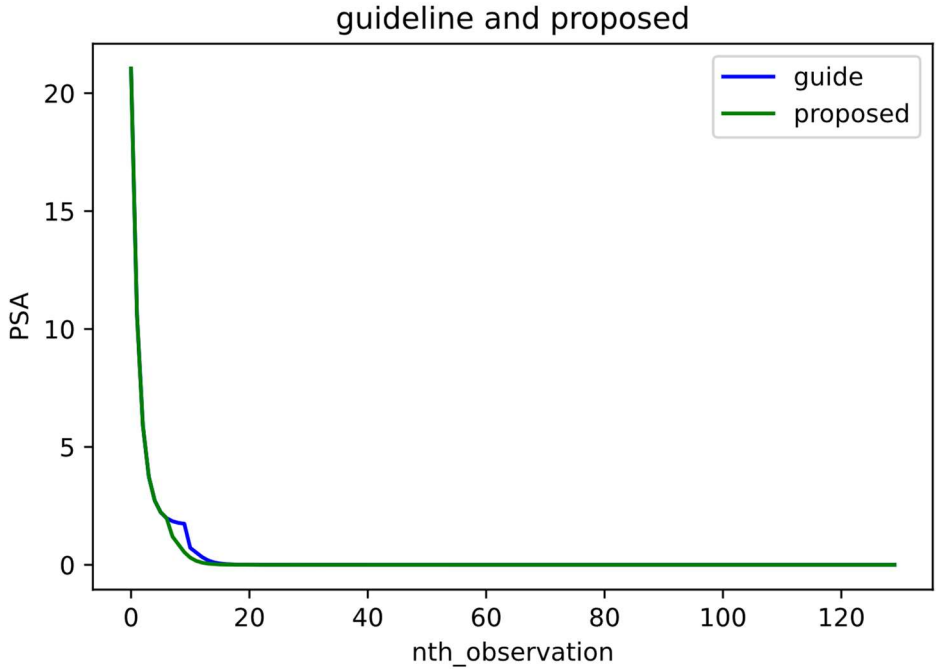

図 11. ガイドラインで間欠投薬が中止にならなかった患者の例 2

次に全 130 回の PSA 値の観測のうち最初の 10 回を除いた残りの 120 回の観測の PSA 最大値の分布を比較した(図 12). ガイドラインに従って投薬した場合の PSA 最大値の分布は 10~14 あたりに集まっており, 提案手法に従って投薬した場合の PSA 最大値の分布は 0~5 あたりに集まっている. また, 25%, 50%, 75%, 90%点についてどれも提案手法の方が数値は低いことが分かる（表 3). したがって, 提案手法の方が PSA の値をよく抑えられており, すなわち腫瘍増殖をより効果的に抑えられていることが読み取れる.

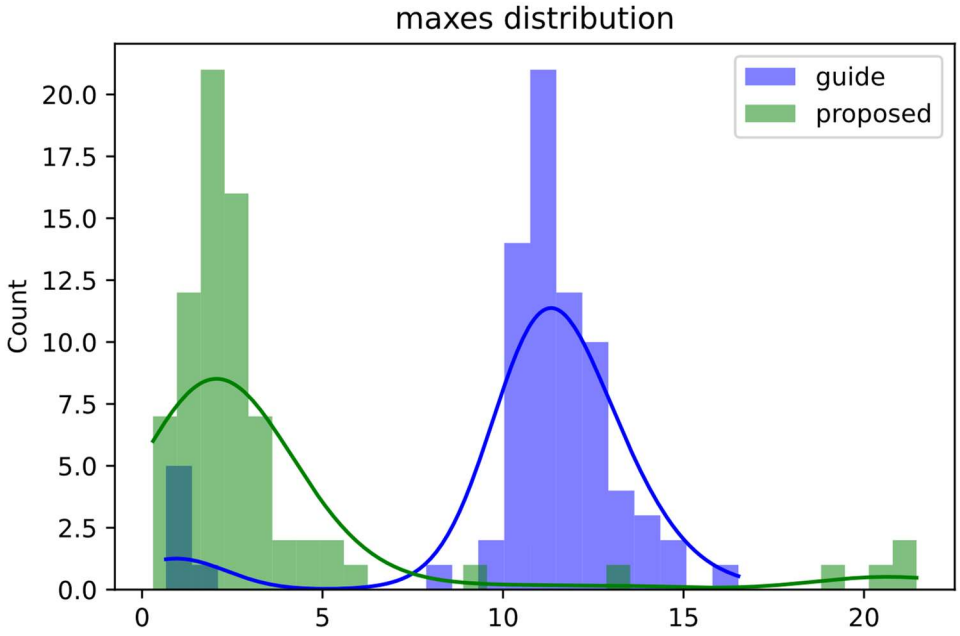

図 12. ガイドラインに従って間欠投薬した場合(guide)と
提案手法に従って間欠投薬した場合(proposed)の最大値分布比較

表 3. ガイドラインに従って投薬，提案手法に従って投薬
それぞれの場合における PSA 最大値の 25%, 50%, 75%, 90%点

	25%	50%	75%	90%
ガイドライン	10.69	11.22	12.23	13.27
提案手法(オフライン)	1.63	2.17	3.04	5.18

次に学習データの数を 100 人分, 200 人分, 400 人分と絞った場合の結果を示す(図 13, 14, 15). 図 12 と同じく全 130 回の PSA 値の観測のうち最初の 10 回を除いた残りの 120 回の観測の PSA 最大値の分布を比較した．図 13, 14, 15 より，どの場合においても PSA 最大値が 7~17 付近に集まっており，ガイドラインに従って投薬した場合の PSA 最大値の分布は 10~14 あたりに集まっている．したがって，提案手法の方は分散が大きく，ガイドラインと比べて腫瘍増殖を効果的に抑制できている場合もあり，抑制できていない場合も多くみられる．

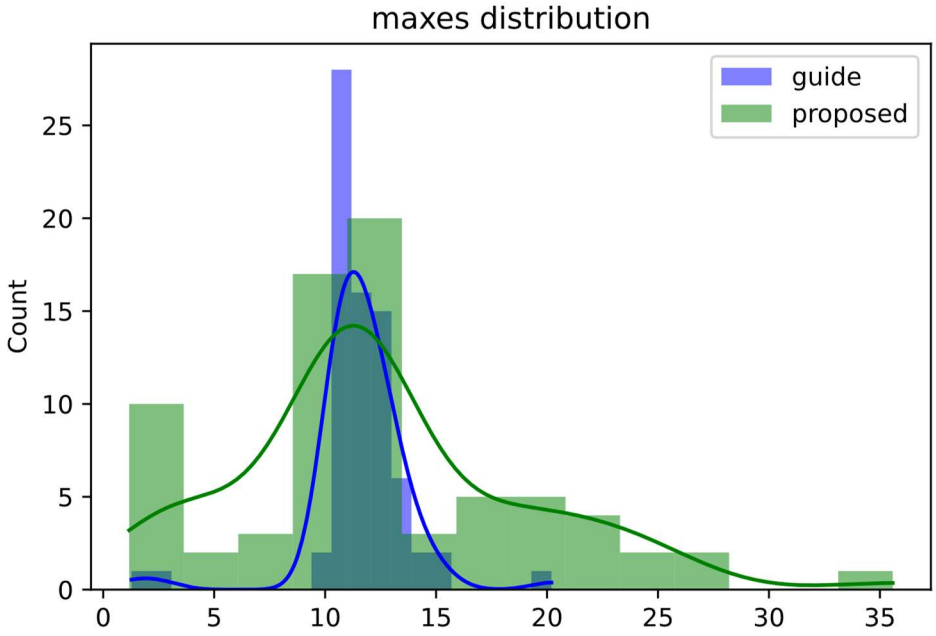

図 13. 学習データを 100 人分とした場合の最大値分布比較

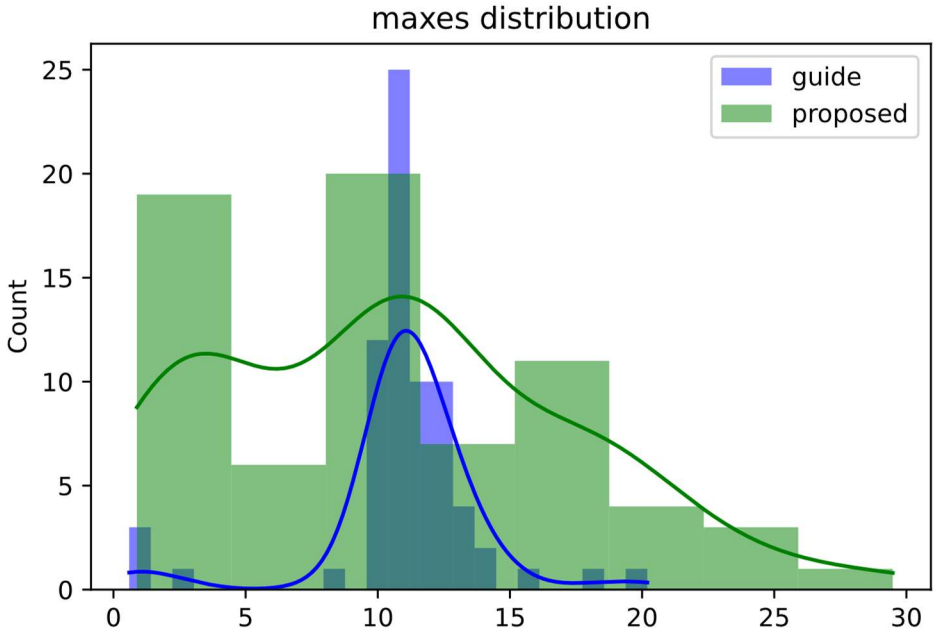

図 14. 学習データを 200 人分とした場合の最大値分布比較

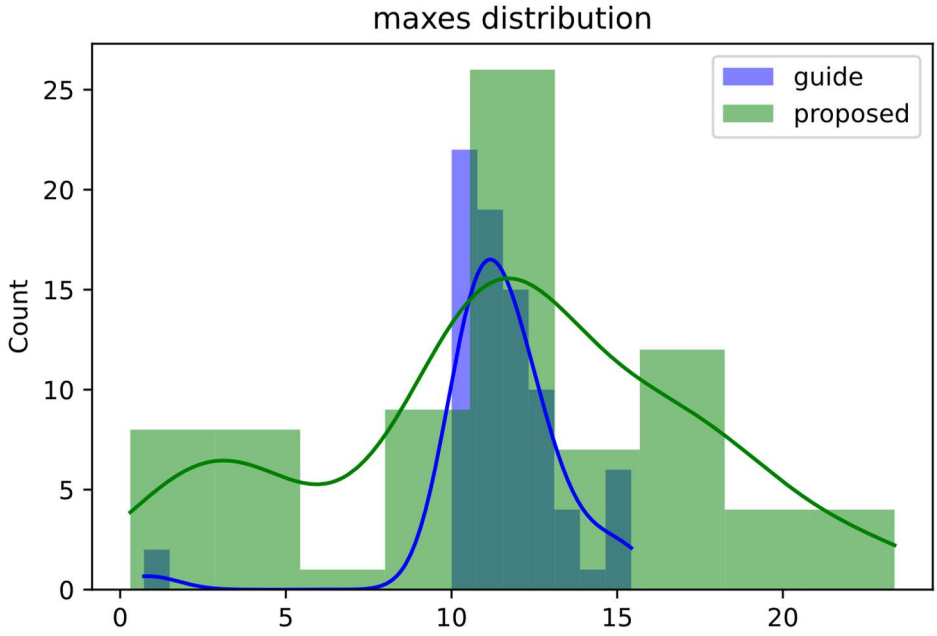

図 15. 学習データを 400 人分とした場合の最大値分布比較

6.2 オンライン Q 学習の結果

テストデータ 100 人の擬似患者に対して, 6.1 のオフライン学習によって得られた Q ネットワークを患者 1 人のデータが得られるごとにモデルの更新を行いながら適用した場合と，ガイドラインに従って間欠的な投薬を行った場合のそれぞれについて, 10 年分の PSA の推移結果を得た．ガイドラインで試験期間満了前に間欠投薬を中止と判断された患者とそうでない患者の例をそれぞれ図 16, 17 と図 18, 19 に示す．横軸が PSA 値の観測回数(28 日ごとに 1 回観測)，縦軸が PSA 観測値を表す．図 16, 17 はガイドラインに従って間欠投薬中止と判断され，その後投薬を継続した患者の例で，図 16 の例から提案手法(proposed)と比べてガイドライン(guide)の方は最終的に PSA 値が再燃してしまっていることが分かる．図 17 の例は提案手法もガイドラインも最終的に PSA 値が再燃してしまっているが，提案手法では PSA 値の増加がガイドラインに比べて遅いことが分かる．図 18, 19 はガイドラインに従って間欠投薬が最後まで続いた患者例で，図 18 の例は提案手法の方が PSA を低く抑えられていることが分かる．図 19 の例は提案手法もガイドラインもほぼ同じ程度に PSA を低く抑えられていることが分かる．

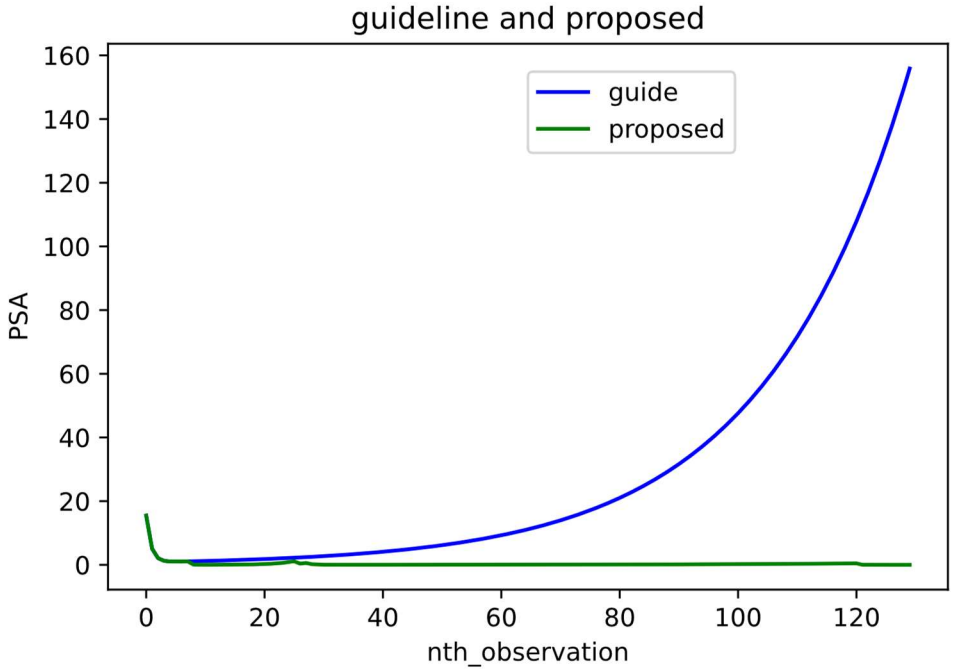

図 16. ガイドラインで間欠投薬が中止になった後投薬を継続した患者の例 1

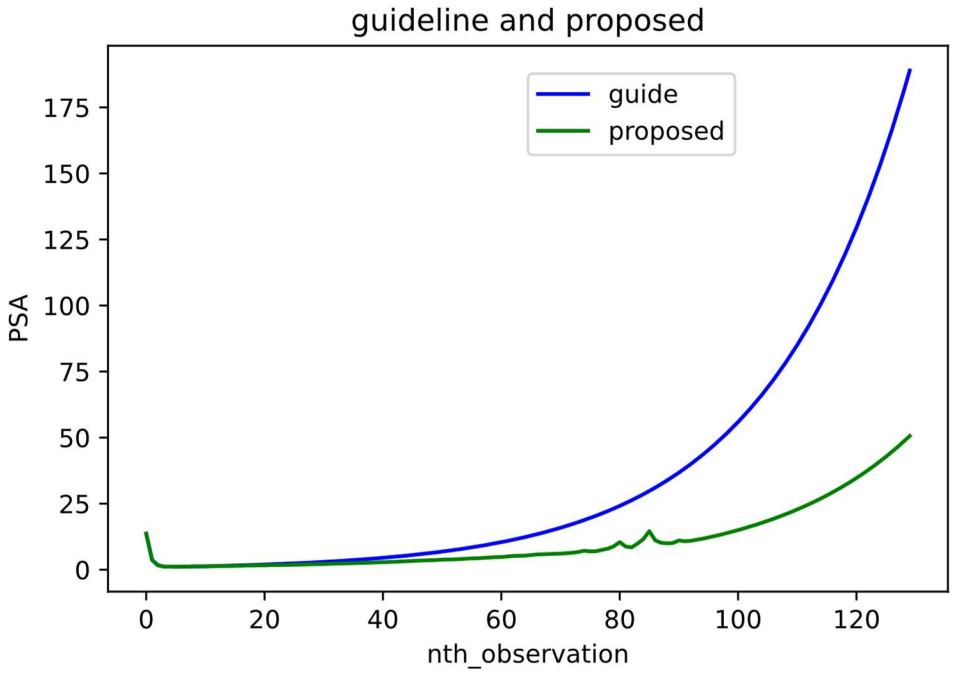

図 17. ガイドラインで間欠投薬が中止になった後投薬を継続した患者の例 2

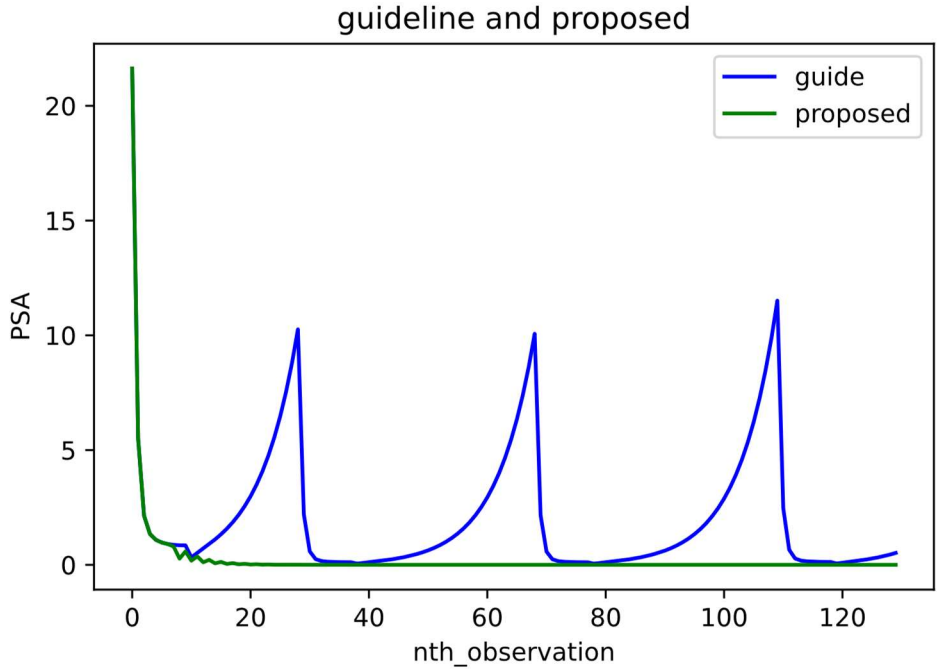

図 18. ガイドラインで間欠投薬が中止にならなかった患者の例 1

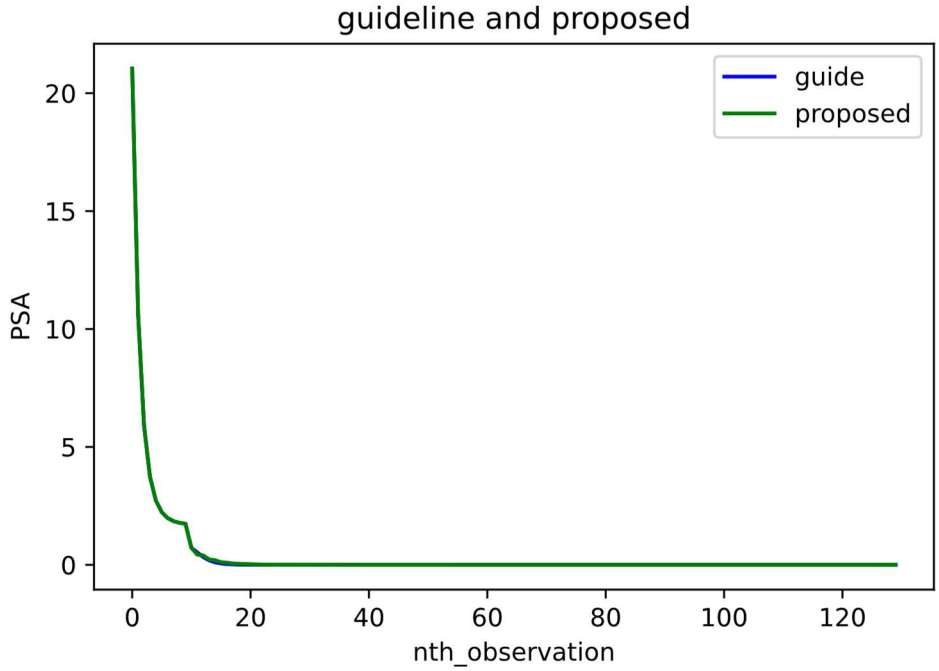

図 19. ガイドラインで間欠投薬が中止にならなかった患者の例 2

次に全 130 回の PSA 値の観測のうち最初の 10 回を除いた残りの 120 回の観測の PSA 最大値の分布を比較した(図 20). ガイドラインに従って投薬した場合の PSA 最大値の分布は 10~14 あたりに集まっており, 提案手法に従って投薬した場合の PSA 最大値の分布は 0~1.5 あたりに集まっている. また, 25%, 50%, 75%, 90%点についてどれもオンライン Q 学習が最も数値は低いことが分かる (表 4).

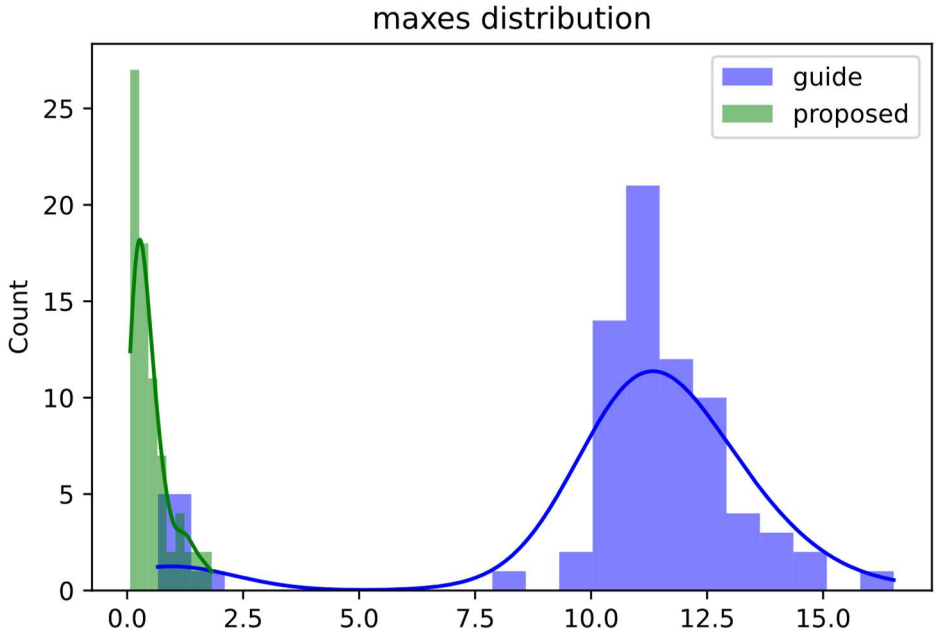

図 20. ガイドラインに従って間欠投薬した場合(guide)と提案手法に従って間欠投薬した場合(proposed)の最大値分布比較, 外れ値削除済み(100 サンプル中 1 サンプル)

表 4. ガイドラインに従って投薬, 提案手法に従って投薬
それぞれの場合における PSA 最大値の 25%, 50%, 75%, 90%点

	25%	50%	75%	90%
ガイドライン	10.69	11.22	12.23	13.27
提案手法(オフライン)	1.63	2.17	3.04	5.18
提案手法(オンライン)	0.19	0.37	0.66	1.21

次に学習データの数を 100 人分, 200 人分, 400 人分と絞った場合の結果を示す(図 21, 22, 23). 図 20 と同じく全 130 回の PSA 値の観測のうち最初の 10 回を除いた残りの 120 回の観測の PSA 最大値の分布を比較した. 図 21, 22, 23 より, どの場合においても PSA 最大値が 0~1.5 付近に集まっており, ガイドラインに従って投薬した場合の PSA 最大値の分布は 10~14 あたりに集まっている. したがって, 学習データの数が少なくても, 提案手法(オンライン)の方が PSA の値をよく抑えられており, すなわち腫瘍増殖をより効果的に抑えられていることが読み取れる.

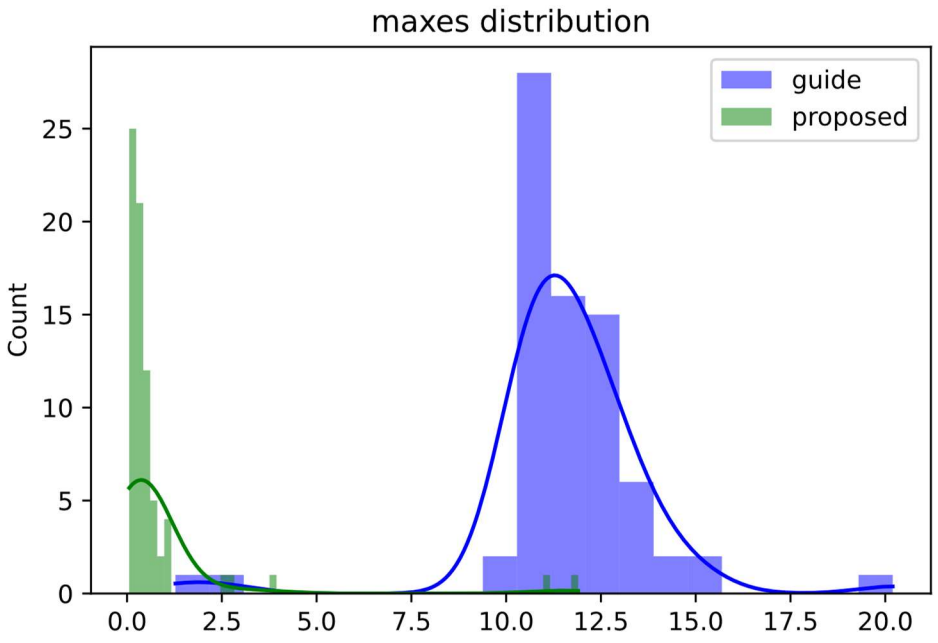

図 21. 学習データを 100 人分とした場合の最大値分布比較

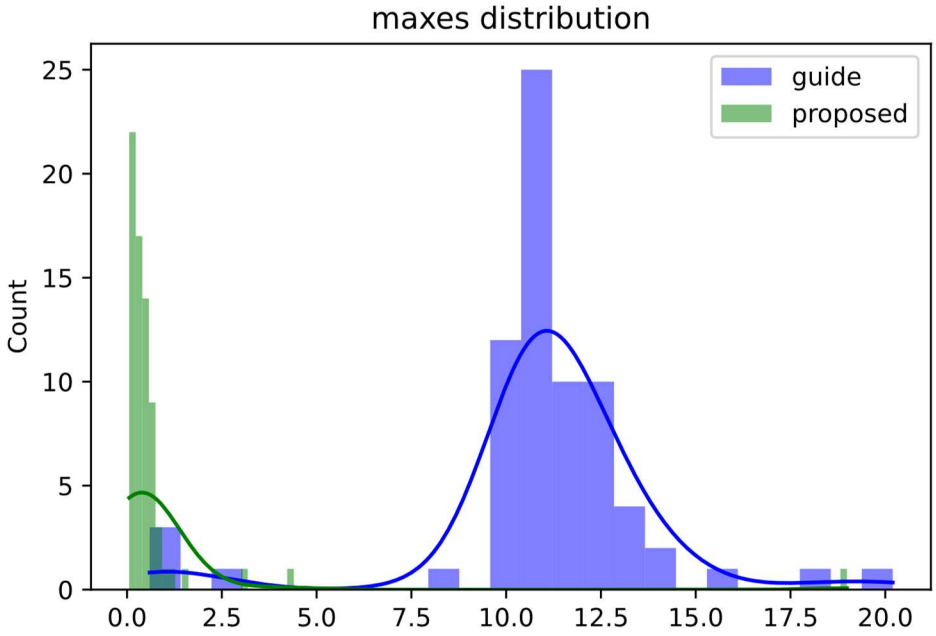

図 22. 学習データを 200 人分とした場合の最大値分布比較

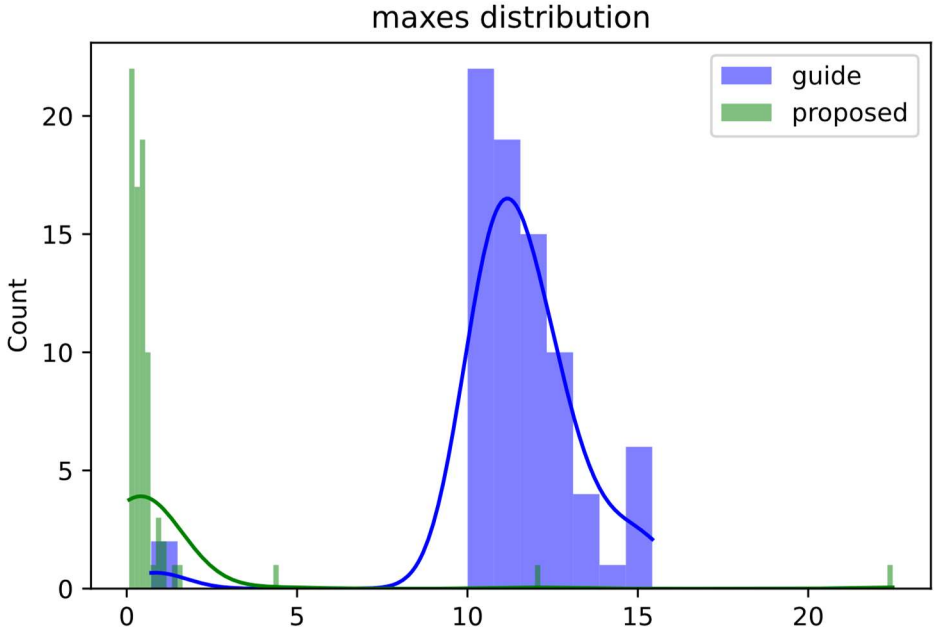

図 23. 学習データを 400 人分とした場合の最大値分布比較

7 考察

シミュレーションデータを用いて前立腺がん細胞の増殖を抑える実験を行った結果，深層強化学習に基づく投薬スケジューリングは，先行研究のガイドラインに比べて増殖を効果的に抑制できることが示された．これは，ガイドラインでは基本的に過去 1 回分の PSA 値のみを参考に投薬の意思決定を行っているのに対し，提案手法では過去 5 回分の PSA 値と 4 回分の投薬状況を参考に投薬の意思決定を行っているため，意思決定に利用できる情報の差によるものと考えられる．また，提案手法では報酬の設定により，将来における PSA 値をできる限り小さくするように学習しており，学習した Q ネットワークに従って行う意思決定は，ガイドラインのようなシンプルな場合分けよりも強力といえる．

また，全体的にオンライン Q 学習はオフライン Q 学習に比べて増殖を効果的に抑制できることが示された．これは，オフライン Q 学習ではガイドラインに従って得られたデータ(学習データ)のみを用いて学習し，その後 Q ネットワークの更新をしなかった一方，オンライン Q 学習ではガイドラインに従って得られたデータに加え，DQN に従って得られたデータを収集し Q ネットワークの更新を続けたことによるものと考えられる．

学習を行うために必要なデータ数を確認するため，学習データの数が 100 人分, 200 人分, 400 人分，それぞれの場合においてテストデータに対して PSA 抑制の程度を検証した(図 13, 14, 15, 21, 22, 23)．オフライン Q 学習の結果(図 13, 14, 15)ではどの場合においても学習データが 835 人分の場合(図 12)より性能が劣っていたため，本研究においては学習データを 835 人分程度若しくはそれ以上にすることで良い結果が得られる可能性が高いと考えられる．オンライン Q 学習の結果(図 21, 22, 23)ではどの場合においても 835 人分の場合(図 20)と同様な結果が得られたため，学習データの数が少なくても済む可能性が高い．

深層強化学習には，意思決定の根拠の説明ができないという課題がある．医療のように人命の関わる意思決定においては，意思決定の誤りが取り返しのつかない事態を引き起こす可能性があるため，判断根拠の明確でないブラックボックスモデルの出力をそのまま採用することには大きなリスクがある．そこで，ガイドラインの方策の一部を導入すること，例えば PSA が 10 を超えると強制的に投薬に切り替えるようにすることによって保守的な意思決定を行うことが考えられる．PSA が 10 以下である場合は深層強化学習の方策に従って行動し PSA を効果的に抑え, 10 を超える場合は深層強化学習モデルの想定外の出力によって PSA が急激に増加しないようにガイドラインの方策で PSA の上限を抑える．このように深層強化学習モデルにすべての意思決定をゆだねるのではなく，あくまでも医師による意思決定の支援を行うものとして導入することが有効と考えられる．

8 結論

本研究では，前立腺がんの間欠的な治療における腫瘍増殖率を抑制するという課題を解決するため，患者一人ひとりに対して最適な治療を選択するという動的治療計画の考え方を取り入れ，深層強化学習という手法で課題解決を行った．その結果，先行研究のガイドラインよりも腫瘍の増殖を効果的に抑えられる方策が得られた．また，オフライン Q 学習で学習した Q ネットワークを更新せずに用いるよりも，オンライン Q 学習でモデルを更新し続けた方が良いということが分かった．学習に必要なデータ数はオンライン学習の場合，最初は 100 人分程度でもオンラインで Q ネットワークを継続的に更新すれば最終的に良いモデルが得られる．ただし，最初の人数が多ければ多いほどオフライン学習の時点で良い結果が得られる可能性が高いため，最初はできるだけ多くのデータを用いてオフライン Q 学習を行い，その後オンラインで Q ネットワークを更新し続けるのが良いと考えられる．

オフライン Q 学習を行う際に，分布シフトのようなオフライン強化学習特有の問題が潜在している可能性があるため，今後モデルの改良が重要になると考えられる．今回は情報として PSA 値と投薬状況のみを用いたが，性別，年齢，投薬量，薬の種類などの変数を追加してモデル構築，または CQL, RNN のようなモデルに変更することでモデルが改良できると考えられる．

ニューラルネットワークの説明可能性の問題については「モデルにしたがって意思決定を行うが, PSA が一定値を超えるとガイドラインの方策に切り替えるもしくは継続投薬に切り替える」といったようにモデルの応用に制限をかけることによって問題の重さが緩和できると考えられる．したがって，実際の医療現場を想定すると，医師の経験と勘と機械学習モデルによる意思決定の支援の組み合わせにより間欠的な治療のプロトコルが確立される可能性がある．

また，モデルを仮定せずとも前立腺がんの腫瘍増殖を効果的に抑えることができたため前立腺がん以外で再燃をするタイプの課題に対しても同じように深層強化学習を用いてモデル構築をすることで意思決定の支援が期待できる．

謝辞

本論文の執筆にあたり，主指導教員である岩山幸治准教授には，研究手法から論文執筆までたくさんご指導をいただき感謝の意を表します．課題からずれた時に軌道修正に関するアドバイスをいただいたので研究がとてもスムーズに進められました．

また，滋賀大学の先生方の講義やゼミ生の皆様の研究がハイレベルで良い刺激を受けましたことに深く感謝申し上げます．

最後に，本論文の執筆にあたり，心身ともに支えてくれた両親に感謝いたします．

引用・参考文献

[1] World Health Organization. Globocan 2020. http://globocan.iarc.fr : accessed on August 9, 2022

[2] What's? 前立腺がん https://www.zenritsusen.jp/ : accessed on August9, 2022

[3] 日本泌尿器科学会 https://www.urol.or.jp/public/symptom/08.html: accessed on August 9, 2022

[4] がん情報サービス https://ganjoho.jp/public : accessed on August9, 2022

[5] Niraula S, Le LW, Tannock IF: Treatment of prostate cancer with intermittent versus continuous androgen deprivation: a systematic review of randomized trials. Journal of Clinical Oncology. 2013

[6]前立腺癌治療ガイドライン，日本泌尿器科学会. 2016: https://www.urol.or.jp/lib/files/other/guideline/23_prostatic_cancer_2016.pdf

[7] Hirata Y, Bruchovsky N, Aihara K: Development of a mathematical model that predicts the outcome of hormone therapy for prostate cancer. Journal of Theoretical Biology. 2010

[8] Bruchovsky N, Klotz L, Crook J, Malone S, Ludgate C, Morris WJ, Gleave ME, Goldenberg SL: Final results of the Canadian prospective phase II trial of intermittent androgen suppression for men in biochemical recurrence after radiotherapy for locally advanced prostate cancer: clinical parameters. Cancer. 2006

[9] Chakraborty B, Moodie E: Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine. New York: Springer. 2013

[10] Parmigiani G: Modeling in medical decision making: A Bayesian approach. New York: Wiley. 2002

[11] Diederik P, Jimmy B: Adam: A method for stochastic optimization: The International Conference on Learning Representations. 2015

[12] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16. 2002