Reinforcement learning and Bayesian outcome-weighted learning for precision medicine : integrating medical knowledge into decision-making algorithms

Sophia Yazzourh

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Apprentissage par renforcement et outcome-weighted learning bayésien pour la médecine de précision. Intégration de connaissances médicales dans les algorithmes de décision.

Thèse présentée et soutenue, le 22 octobre 2024 par

Sophia YAZZOURH

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Mathématiques et Applications

Unité de recherche
IMT : Institut de Mathématiques de Toulouse

Thèse dirigée par
Nicolas SAVY et Philippe SAINT PIERRE

Composition du jury
M. Raphaël PORCHER, Président, Université Paris Cité
M. Rodolphe THIEBAUT, Rapporteur, Université de Bordeaux
Mme Julie JOSSE, Rapporteure, INRIA Antenne de Montpellier
Mme Emmanuelle CLAEYS, Examinatrice, Université Toulouse III - Paul Sabatier
M. Nicolas SAVY, Directeur de thèse, Université Toulouse - Jean Jaurès
M. Philippe SAINT-PIERRE, Co-directeur de thèse, Université Toulouse III - Paul Sabatier

À mes parents, Monique et Sabr

Remerciements

Je tiens tout d'abord à exprimer ma plus profonde gratitude à mes directeurs de thèse, Nicolas et Philippe. Merci pour vos précieux conseils, pour votre accompagnement attentif, et pour avoir contribué par renforcement à mon développement personnel et professionnel. Votre bienveillance et votre engagement ont grandement enrichi mon parcours. Travailler à vos côtés a été un immense plaisir.

I would also like to extend my heartfelt thanks to Nikki Freeman, whose valuable advice and sisterhood made Bayesian research much easier. Additionally, I would like to thank Professor Kosorok for his warm welcome to his lab and for the valuable time he dedicated to my research.

Ce sont mes parents que je voudrais le plus remercier pour leur soutien indéfectible. Vous êtes incroyables, et je vous suis profondément reconnaissant pour tout ce que vous avez fait. Je tiens également à remercier chaque membre des familles Martin, Meunier et Yazzourh pour leur affection. J'embrasse particulièrement mes deux cousines, Clara et Fati, qui sont mes deux rayons de soleil et mes plus grandes fiertés.

Julie et Heloïse, vous êtes les meilleures amies qu'on puisse rêver d'avoir et vous me manquez quotidiennement. Merci pour ces saisons et merci d'avance pour ces prochaines années.

Je voudrais remercier "procédure" pour le bureau 206, car c'est là que j'ai fait mes plus belles rencontres. Nicolas, juste en face, tu as été mon plus fidèle complice durant ces trois ans, et tu as le plus beau rire du monde. Candice, merci de m'avoir rejoint ; tu as rendu chaque minute du séminaire étudiant agréable. Tu es une pédagogue et une amie incroyable. Benjamin, je te remercie pour ton affection, tes goûts musicaux, ta cuisine et tes imitations. Florian, merci pour ton crâne, ton humour et ta générosité sans faille. Je tiens également à remercier les membres honoraires du bureau 206 : Niels, Ana et Adama. J'espère que celle ou celui qui prendra ma place se rendra compte de la chance qu'elle ou il a.

J'aimerais faire une dédicace aux doctorants et guests de ma promotion : Étienne, Anthony M., Armand, Anthony SC., Fanny, Lucas C., Mitjia, Erwanne et Pauline. Merci pour tous les moments que nous avons partagés. Je garde de précieux souvenirs avec chacun d'entre vous, que ce soit autour d'un café ou d'une bière. J'embrasse aussi à ce moment Joachim et Alexandre qui ont suivi et partagé nos aventures.

Merci aux anciens du laboratoire d'avoir été une source d'inspiration et d'avoir toujours donné d'excellents conseils. Laetitia, Michèle, Perla, Maxime, Viviana, Javier, Luca, Alberto, Clément B., Hippolyte et Fushi, je suis admirative de voir ce que vous êtes devenus. Bonne chance à ceux qui restent encore un peu : Elio, Louis C., Louis D., Paul, Niki, Arnaud et Angel. Vous êtes géniaux et vous avez une énergie solaire.

Je voudrais également remercier l'INSA pour les rencontres que j'y ai faites. May, merci pour ton soutien indéfectible. À ma dream team : Paul, Camille, Emmeline, Oumaima et Louis, merci pour toutes les aventures que nous avons vécues ensemble et celles qui sont encore à venir.

Merci à Camille et Blandine pour leurs histoires, aussi palpitantes que nos soirées.

Je voudrais remercier le handball pour avoir été une véritable source de passion et d'émotions, et pour m'avoir permis de rencontrer des personnes incroyables. Lou, Léa, Audrey, Marie, Naé, Julie C., Solène et, de manière générale, toutes les filles du CRAHB, du TFH et de l'INSA, merci pour ces moments passés sur et en dehors du terrain.

Merci Toulouse pour ces 10 dernières années. Merci l'Institut de Mathématiques de Toulouse pour mes plus belles années d'études. Adieu l'UPSIDUM.

Contents

1 Introduction en français **1**
- 1.1 Médecine de précision . . . 1
  - 1.1.1 Généralités . . . 1
  - 1.1.2 *Individualized treatment regime* . . . 2
  - 1.1.3 *Dynamic treatment regimes* . . . 2
  - 1.1.4 Méthodes de constructions de règles de décision . . . 3
- 1.2 Apprentissage par renforcement . . . 3
  - 1.2.1 Qu'est que l'apprentissage par renforcement ? . . . 3
  - 1.2.2 Apprentissage par renforcement et médecine de précision . . . 5
  - 1.2.3 Intégration du savoir médical dans les modèles d'apprentissage par renforcement . . . 6
  - 1.2.4 Construction de récompenses par apprentissage par préférences . . . 7
- 1.3 *Outcome-Weighted Learning* . . . 8
  - 1.3.1 Une méthode de classification pondérée . . . 8
  - 1.3.2 Quantification d'incertitude et *Bayesian OWL* . . . 8
- 1.4 Organisation du manuscrit . . . 9

2 Introduction **11**
- 2.1 Precision Medicine . . . 11
  - 2.1.1 Overview . . . 11
  - 2.1.2 Individualized treatment regime . . . 12
  - 2.1.3 Dynamic treatment regimes . . . 12
  - 2.1.4 Decision rule construction methods . . . 12
- 2.2 Reinforcement learning . . . 13
  - 2.2.1 What is reinforcement learning? . . . 13
  - 2.2.2 Reinforcement learning and precision medicine . . . 14
  - 2.2.3 Integrating medical knowledge into reinforcement learning models . . . 15
  - 2.2.4 Reward construction through preference learning . . . 16
- 2.3 Outcome-Weighted Learning . . . 17
  - 2.3.1 A weighted classification method . . . 17
  - 2.3.2 Uncertainty quantification and Bayesian OWL . . . 17
- 2.4 Structure of the manuscript . . . 18

3 Reinforcement learning for dynamic treatment regimes **21**
- 3.1 Introduction . . . 21
- 3.2 Theoretical foundations of reinforcement learning . . . 23
  - 3.2.1 Decision process . . . 23
  - 3.2.2 Policy . . . 25
  - 3.2.3 Rewards, valuation and optimization of policies . . . 26

3.2.4 Reinforcement learning . . . 30
3.3 The multi-decision setting : dynamic treatment regimes . . . 33
3.3.1 Dynamic treatment regimes . . . 33
3.3.2 Decision process and dynamic treatment regimes . . . 33
3.3.3 Specificities of the medical context . . . 34
3.3.4 Real data application . . . 35
3.3.5 Properties of RL applied to DTR . . . 36
3.4 The single-decision setting : individualized treatment regime . . . 40
3.4.1 Individualized treatment regime . . . 40
3.4.2 Decision process and individualized treatment regime . . . 41
3.4.3 Causality . . . 42
3.5 Conclusion . . . 44

4 Integrating medical knowledge into RL models 45
4.1 Introduction . . . 45
4.2 Approaches to integrating medical knowledge into RL . . . 47
4.2.1 Medical knowledge and model preparation . . . 47
4.2.2 Medical knowledge and rewards . . . 48
4.2.3 Medical knowledge and value functions . . . 49
4.2.4 Medical knowledge and objective function . . . 51
4.2.5 Medical knowledge and policy . . . 51
4.3 Rewards construction based on preference learning . . . 52
4.3.1 Preference learning . . . 53
4.3.2 BMI data application . . . 55
4.3.3 Cancer application . . . 61
4.3.4 Conclusion . . . 66
4.4 Perspectives . . . 68

5 Bayesian OWL 73
5.1 Introduction . . . 74
5.2 Background . . . 76
5.2.1 Setting . . . 76
5.2.2 Outcome-weighted learning . . . 76
5.2.3 Bayesian support vector machines . . . 77
5.3 Our approach . . . 78
5.3.1 Prior specification for the ITR parameters . . . 79
5.3.2 Exponential power prior distribution for $\boldsymbol{\beta}$ . . . 80
5.3.3 Spike-and-slab prior distribution for $\boldsymbol{\beta}$ . . . 81
5.3.4 Estimation . . . 81
5.3.5 Prediction and uncertainty quantification . . . 86
5.4 Simulation studies . . . 86
5.4.1 Classification performance . . . 86
5.4.2 Treatment recommendation uncertainty quantification . . . 88
5.5 Discussion . . . 89

5.6 Appendix : derivation of the Gibbs sampling algorithms 92
5.6.1 Conditional distribution of $\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, \mathbf{a}_i, \mathbf{r}_i$ 92
5.6.2 Conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\mu}_0, \sigma_0^2, \mathbf{r}, \mathbf{a}, \mathbf{x}$ (Normal prior) . . . 94
5.6.3 Conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ (Exponential power prior) 99

6 Conclusion and perspectives **101**

Appendix **107**

Chapitre 1

Introduction en français

1.1 Médecine de précision

1.1.1 Généralités

La recherche médicale permet de mieux comprendre les mécanismes biologiques qui influencent l'évolution et le développement des maladies chroniques, des mécanismes qui varient considérablement d'un patient à l'autre. Grâce aux avancées remarquables de la médecine moderne en termes de soins, de médicaments et de traitements, l'exploration de nouvelles approches se concentre de plus en plus sur l'idée de fournir le bon traitement à la bonne personne, au bon moment. C'est le paradigme de la médecine de précision, également connue sous le nom de médecine personnalisée, qui permet d'adapter les traitements aux caractéristiques individuelles des patients [11, 50, 51]. Son objectif n'est pas de remplacer les traitements existants ou de déterminer un médicament unique pour chaque patient, mais de compléter l'arsenal thérapeutique actuel pour permettre une prise de décision médicale personnalisée, dans le but de soigner chacun de manière efficace. Cela est rendu possible grâce aux avancées technologiques dans la collecte et le stockage des données. Le volume de données individuelles collectées a considérablement augmenté, permettant ainsi de mieux comprendre les facteurs individuels influençant les effets d'une intervention. Ces nouvelles et vastes bases de données permettent de contrôler l'hétérogénéité des patients. Chaque individu est unique, que ce soit en termes de génétique, d'environnement, de mode de vie et particulièrement dans sa réponse aux traitements. Ces variabilités, souvent mises en lumière par les essais cliniques randomisés contrôlés, soulignent l'importance de mettre en place des traitements plus personnalisés, améliorant ainsi la qualité des soins proposés.

Nous nous intéressons ici plus particulièrement aux maladies chroniques telles que le cancer, le diabète ou les troubles psychiatriques... Les patients atteints de ces maladies suivent des traitements de longue durée, régulièrement évalués ou réévalués par les médecins. La médecine de précision se matérialise alors sous la forme de règles de décision qui recommandent les traitements à entreprendre en fonction de l'état du patient. Ces ensembles de règles de décision médicale sont formalisés à travers des régimes de traitements dynamiques, ou *Dynamic Treatment Regimes* (DTR) en anglais. Un DTR consiste en une séquence de règles de décision, une par étape d'intervention, qui dicte comment individualiser les traitements en fonction de l'évolution de l'historique des traitements et des covariables [11]. Lorsqu'il n'y a qu'une seule étape d'intervention médicale, un seul temps de décision, on parle alors de régime de traitement individualisé ou *Individualized Treatment Regime* (ITR) en anglais.

La recherche de règles de décision médicale personnalisées, en particulier dans le cadre des DTR, s'appuie sur des données observationnelles médicales structurées de manière longitudinale. L'analyse de ces données réelles en médecine implique nécessairement de prendre en compte la causalité. En médecine de précision, il est crucial de comprendre les relations causales pour déterminer comment les interventions médicales affectent individuellement les patients. Contrairement aux simples corrélations, les relations causales permettent d'identifier les effets directs des traitements sur la santé, ce qui aide les cliniciens à adapter les traitements aux caractéristiques spécifiques des patients, garantissant ainsi des interventions appropriées et efficaces. Pour tirer des conclusions causales à partir des données observées, il est essentiel de respecter certaines hypothèses, que nous détaillerons dans la Section 3.4.3 du Chapitre 3. L'un des designs d'essais cliniques qui satisfait ces conditions est celui des *Sequential Multiple Assignment Randomized Trials* (SMART). Les essais SMART sont des études expérimentales dans lesquelles les patients sont randomisés à plusieurs moments clés du traitement, afin de tester et d'optimiser des stratégies de traitement adaptatif. Considérés comme le "gold standard" des essais cliniques en médecine de précision, ces designs sont largement étudiés dans la littérature [13, 51, 11]. Cependant, leur mise en œuvre est complexe et coûteuse, ce qui limite le nombre d'essais disponibles. Parmi les essais notables, on peut citer CATIE [107], ADHD [11, 54], et STAR*D [11, 53].

1.1.2 *Individualized treatment regime*

Un premier cadre formel de la prise de décision dans le cadre de la médecine de précision est celui des ITR. On considère un seul point de décision à partir de données observées, de taille $n \in \mathbb{N}$, de la forme $\{(X_i, A_i, R_i)\}_{i=1}^n$, où $X \in \mathcal{X}$ représente les caractéristiques initiales du patient, $A \in \mathcal{A}$ est le traitement administré, et $R \in \mathbb{R}$ est la réponse au traitement, des valeurs plus élevées indiquant un meilleur état du patient. Un ITR est une fonction $d : \mathcal{X} \to \mathcal{A}$. Sous le régime d, les patients avec $X = x$ se verraient attribuer le traitement $d(x)$ [50]. Identifier un ITR optimal, implique de trouver la règle de décision qui maximise la réponse au traitement R pour tous les patients.

1.1.3 *Dynamic treatment regimes*

Le second cadre formel, qui généralise les ITR dans un contexte à plusieurs étapes de décisions, est celui des DTR. Les données se présentent alors sous la forme $\{(X_{1,i}, A_{1,i}, R_{1,i}, \ldots, X_{T,i}, A_{T,i}, R_{T,i})\}_{i=1}^n$ qui comprennent n données identiquement indépendamment distribuées répliques de $(X_1, A_1, R_1, \ldots, X_T, A_T, R_T)$ où $X_1 \in \mathcal{X}_1$ sont les caractéristiques du patient au début de l'étude et $X_t \in \mathcal{X}_t$ sont les données récoltées au temps intermédiaire t pour $t = 2, \ldots, T$; $A_t \in \mathcal{A}_t$ est le traitement assigné au temps t ; R_t est la réponse mesuré au temps t. On introduit également $H_1 = X_1$ et $H_t = (H_{t-1}, A_{t-1}, R_{t-1}, X_t)$ où H_t représente l'historique médical du patient au temps t. Un DTR est une séquence de fonctions $d = (d_1, \ldots, d_T)$ où $d_t : \mathcal{H}_t \to \mathcal{A}_t$ [50], signifiant qu'à chaque étape de décision, un traitement est recommandé en fonction

de l'historique de traitement. Un DTR optimal maximise l'espérance d'une mesure de réponse cumulative du vecteur $\mathbf{R} = (R_1, \ldots, R_T)$.

1.1.4 Méthodes de constructions de règles de décision

Pour construire ces règles de décision optimales, deux grandes familles d'approches existent. La première famille repose sur des modèles de régression tels que le *Q-learning* [126, 75] et la *G-estimation* [96, 99, 95]. La deuxième famille de méthodes inclut des approches qui estiment la valeur attendue de la réponse sous un régime particulier sans faire d'hypothèses paramétriques. Une fois cette valeur estimée, la règle de décision optimale peut être déterminée en explorant une classe de régimes possibles. Parmi ces méthodes, on trouve les *marginal structural mean models* [97, 83], *inverse probability of treatment weighting* [97], *augmented inverse probability of treatment weighting* [139], et *Outcome-Weighted Learning* (OWL) [143].

Pendant cette thèse, notre recherche s'est appuyée sur une méthode de chaque famille de résolution : le *Q-learning* et OWL. Dans la suite de cette introduction, nous nous concentrerons tout d'abord particulièrement sur le *Q-learning*, dans le cadre DTR, un algorithme faisant partie du vaste domaine de l'apprentissage par renforcement. En effet, nous verrons que ce sous-domaine de l'apprentissage automatique, qui répond aux problèmes de décision à plusieurs étapes, peut s'adapter parfaitement au contexte de la médecine de précision dans la Section 1.2. Cela nous permettra de situer deux de nos travaux de recherche : l'intégration des connaissances médicales dans les modèles d'apprentissage par renforcement et la construction de récompenses dites *data-driven* via l'apprentissage par préférences. Ensuite, nous nous reviendrons plus particulièrement sur l'une des méthodes non paramétriques appliquées dans le cadre des ITR : la méthode OWL dans la Section 1.3. Cela nous conduira à introduire la troisième contribution de cette thèse, qui propose une perspective bayésienne pour cette méthode, permettant de quantifier l'incertitude associée aux traitements recommandés.

1.2 Apprentissage par renforcement

1.2.1 Qu'est que l'apprentissage par renforcement ?

Le *Q-learning*, mentionné parmi les méthodes précédentes, fait partie d'un ensemble plus large d'algorithmes appartenant au domaine de l'apprentissage par renforcement, ou *Reinforcement Learning* (RL). Ce domaine de recherche en apprentissage automatique se concentre sur l'acquisition de stratégies de contrôle pour un système, en exploitant les interactions dynamiques entre ce système et un agent intelligent. L'objectif est de permettre à cet agent, une entité apprenante, de maximiser un objectif prédéfini au fil de ses interactions avec l'environnement. Classiquement, le RL se déroule dans un environnement interactif. L'agent choisit des actions, entraînant des modifications de l'état de l'environnement. En retour, l'environnement fournit des récompenses, des valeurs numériques spécifiques, pouvant être positives ou négatives. Ces récompenses indiquent à quel point les actions de l'agent ont été bénéfiques ou nuisibles par rapport

à l'objectif global de maximisation de la récompense cumulative. Grâce à ce retour d'information, l'agent adapte sa stratégie pour améliorer ses décisions futures et atteindre plus efficacement son objectif. Ainsi, le domaine de l'apprentissage par renforcement introduit les notions suivantes :

- **État** : une représentation de la situation actuelle dans laquelle se trouve l'agent. Les états contiennent les informations nécessaires pour décider des actions futures.
- **Action** : une décision prise par l'agent qui affecte l'environnement. Les actions peuvent être discrètes (comme déplacer un pion sur un plateau) ou continues (comme ajuster la vitesse d'une voiture).
- **Récompense** : un signal reçu par l'agent après avoir effectué une action. La récompense indique à quel point l'action était bénéfique en fonction de l'objectif de l'agent.
- **Politique** : une stratégie ou un ensemble de règles que l'agent utilise pour choisir ses actions en fonction des états. La politique peut être déterministe (une action fixe pour chaque état) ou stochastique (une distribution de probabilités sur les actions).
- **Fonctions valeur** : une estimation de l'utilité ou de la récompense future cumulée qu'un agent peut attendre d'un état ou d'une action donnée. Les fonctions de valeur aident à évaluer les politiques et à décider des meilleures actions.

Les fondements et origines de l'apprentissage par renforcement sont attribués à Richard Bellman dans les années 60 [6, 7, 108] grâce à ses travaux sur la programmation dynamique et l'optimalité, dont les équations forment maintenant un des principes clés de ce domaine. Dans les années 90, l'algorithme du *Q-learning* a été proposé par Chris Watkins, marquant une avancée significative [125]. Parallèlement, Richard Sutton a développé la méthode de l'apprentissage par différence temporelle [112]. En 1998, Andrew Barto et Richard Sutton ont publié une première édition de "Reinforcement Learning : An Introduction" [113] qui résume et compile leurs recherches et développements antérieurs. Ce livre, souvent cité comme une source essentielle pour comprendre l'évolution et les fondements de l'apprentissage par renforcement, pose les bases des concepts de récompense et de politique formalisés dans le cadre des processus de décision de Markov.

Le formalisme mathématique classique et largement adopté par la communauté de l'apprentissage par renforcement est celui des Processus de Décision de Markov ou *Markov Decision Process* (MDP). Un MDP est défini par un quintuplet (S, A, P, R, γ) [9, 28] où :

- **États** : ensemble des états possibles du système, S.
- **Actions** : ensemble des actions disponibles pour l'agent, A.
- **Probabilités de transition** : fonction $P(s'|s, a)$ donnant la probabilité d'atteindre l'état s' après avoir pris l'action a dans l'état s.
- **Récompense** : fonction $R(s, a)$ fournissant la récompense reçue après avoir pris l'action a dans l'état s.
- **Facteur d'actualisation** : paramètre $0 \leq \gamma \leq 1$ qui pondère l'importance des récompenses futures.

L'objectif est de trouver une politique optimale π^* qui maximise la récompense

cumulative attendue à long terme à partir de chaque état s.

L'apprentissage par renforcement trouve des applications variées dans de nombreux domaines grâce à sa capacité à résoudre des problèmes complexes de prise de décision séquentielle. Dans les jeux et simulations, il est utilisé pour former des agents à des jeux complexes comme le Go et les échecs. En robotique, le RL permet aux robots de naviguer, d'éviter des obstacles et de manipuler des objets avec précision. Dans l'automatisation industrielle, il optimise les chaînes de production et la gestion des ressources en temps réel. En finance, il aide à développer des stratégies de trading algorithmique et à optimiser la gestion des risques. Dans le transport et la logistique, le RL est utilisé pour entraîner des véhicules autonomes et optimiser la chaîne d'approvisionnement. Il améliore également la personnalisation du contenu dans les systèmes de recommandation. Dans le cadre de cette thèse, c'est l'application du RL en santé qui nous intéresse, et qui a suscité un engouement dans la recherche médicale à travers de nombreuses revues [133, 16, 21]. Le RL peut servir la santé de nombreuses manières, allant de la logistique et gestion hospitalière aux soins des patients. Ici, nous nous concentrerons particulièrement sur son utilisation en médecine de précision pour l'apprentissage de stratégies de traitements personnalisés.

1.2.2 Apprentissage par renforcement et médecine de précision

L'apprentissage automatique a connu un développement significatif au cours des dernières décennies, offrant des solutions efficaces pour résoudre des problèmes complexes de grande envergure. Parmi ces approches, RL s'est distingué par sa capacité à apprendre des règles de décision dans des scénarios séquentiels, démontrant ainsi son utilité en médecine de précision. Le RL vise principalement à identifier des règles de décision optimales en maximisant les gains cumulatifs à long terme. En médecine, où les effets des traitements et leurs éventuels effets secondaires peuvent se manifester après plusieurs étapes, l'élaboration d'une stratégie à long terme est un atout majeur. De plus, les modèles de RL permettent de traiter simultanément de vastes ensembles de données de covariables des patients tout en abordant des problèmes de décision à plusieurs étapes.

La modélisation du problème de décisions séquentielles sous-jacent établit un lien entre les DTR et le RL. Ainsi, le RL s'est largement imposé parmi les méthodes de détermination des DTR au cours de ces dernières années [127, 19, 121, 15, 60]. Cependant, la transition entre RL et DTR, bien que évidente, n'est pas nécessairement directe. Le RL, tel qu'il est présenté classiquement, n'est pas adaptable aux questions liées à l'application médicale, et certains algorithmes répondent mieux aux attentes de ce domaine. Premièrement, décrire un DTR par un processus de décision markovien peut être limitant à cause de l'hypothèse markovienne. Cette dernière stipule que le dernier état du patient contient toute l'information nécessaire à la prise de décision, obligeant ainsi à ignorer son passé médical. C'est une hypothèse forte pour le cadre de la santé, où il est préférable de considérer un processus de décision plus général. Deuxièmement, le RL est généralement introduit dans un contexte d'environnement interactif. Or, il est évident qu'il est impossible et non éthique de laisser un algorithme

interagir avec des patients pour apprendre par tâtonnement une stratégie de traitement optimale. Cette stratégie de traitement doit être déterminée à partir de données déjà collectées. Nous sommes alors dans un contexte particulier de RL appliqué en mode *offline*. Plus précisément, les données collectées l'ont été selon une stratégie médicale déjà établie, soit par un médecin, soit dictée par un essai clinique. Nous apprenons donc une stratégie optimale par rapport à celle déjà suivie, ce qui est désigné par le terme de RL *off-policy*. Le RL est régi par différentes dichotomies : *model-based/model-free*, *policy-based/value-based* ou encore *on-policy/off-policy*. Ces propriétés peuvent nous aider à nous y retrouver parmi la vaste littérature algorithmique du RL. Cela m'amène à mon troisième point sur les spécificités de l'application du RL aux DTR : le *Q-learning*. Parmi la grande variété d'algorithmes, le *Q-learning* dans sa forme *backward* s'est imposé comme l'algorithme idéal pour l'application aux DTR (voir l'Annexe 6). Les propriétés du *Q-learning*, étudiées en détail, montrent pourquoi cet algorithme est particulièrement adapté aux besoins de la médecine de précision.

Le lien entre les DTR et le RL, ainsi que les éléments soulevés ici, feront l'objet du Chapitre 3, permettant d'approfondir le formalisme du RL, les spécificités du contexte applicatif et les propriétés algorithmiques.

1.2.3 Intégration du savoir médical dans les modèles d'apprentissage par renforcement

Bien que le RL soit une solution technique prometteuse pour les questions de médecine de précision, il repose sur des techniques d'apprentissage automatique qui peuvent susciter de l'appréhension chez les patients comme chez les praticiens. S'assurer que les règles de décision construites sont sûres, interprétables et efficaces sur le plan médical est l'une des problématiques majeures de l'application concrète du RL en milieu hospitalier [21, 133]. Un des moyens de répondre à cette problématique est d'intégrer le savoir médicale dans les modèles de RL. En combinant les connaissances cliniques aux algorithmes d'apprentissage automatique, il est possible de créer des modèles plus robustes et adaptés aux réalités du terrain. L'idée est de créer une synergie entre les capacités d'apprentissage automatique et les connaissances des experts du domaine [39, 69]. Cette collaboration renforce la confiance dans les modèles de RL et leurs recommandations [66], tout en facilitant l'adoption de cette technologie par les professionnels de la santé et les patients en milieu clinique [40]. La combinaison de l'apprentissage automatique et de l'expertise humaine produit des résultats supérieurs à ceux obtenus par l'utilisation seule du RL ou par les seules décisions des experts [5, 56]. En outre, d'un point de vue technique, l'intégration des connaissances médicales réduit le temps d'apprentissage, permettant une adaptation et une amélioration plus rapides des méthodes, ce qui mène à des solutions de santé plus efficaces et centrées sur le patient.

L'intégration des connaissances médicales dans les algorithmes d'apprentissage par renforcement commence par la préparation du modèle, notamment par le traitement et la préparation des données ainsi que par la sélection de l'algorithme le plus approprié. Ensuite, cette intégration intervient particulièrement sur les éléments clés du RL tels que les récompenses, les fonctions de valeur et d'objectif, ainsi que la politique. La

première partie du Chapitre 4 présentera un état de l'art des méthodes d'intégration des connaissances médicales dans les modèles d'apprentissage par renforcement, en mettant l'accent sur les propriétés algorithmiques choisies par chacune de ces méthodes. Cela permettra de mettre en lumière les adaptations nécessaires pour l'application des DTR sur des données observationnelles.

1.2.4 Construction de récompenses par apprentissage par préférences

Dans le contexte de l'intégration du savoir d'expert, l'apprentissage par préférences, ou *Preference Learning*, se révèle être une méthode particulièrement prometteuse. L'idée consiste à construire les récompenses d'un modèle de RL à partir de préférences fournies par un expert. Les récompenses jouent un rôle crucial dans l'apprentissage des stratégies car elles constituent le critère à optimiser. Ainsi, leur conception doit encapsuler au mieux l'état du système pour fournir de réelles indications pendant l'apprentissage. Généralement, les récompenses sont définies par un expert du système qui propose de les évaluer via un score. Par exemple, dans les essais cliniques pour les personnes atteintes d'obésité visant à réduire leur poids, la récompense peut être mesurée par l'indice de masse corporelle [61]. Dans les soins intensifs, les traitements peuvent être évalués en fonction des taux de survie ou de mortalité [101]. Certaines récompenses sont plus complexes et combinent plusieurs variables. Par exemple, dans une simulation de cancer présentée dans [144], les récompenses sont basées sur la taille de la tumeur, la toxicité du traitement, le bien-être du patient et les taux de survie. En cas de décès, un score arbitraire de -60 est généralement attribué. Construire manuellement une fonction de récompense peut impliquer des choix arbitraires ou très spécifiques au contexte, ce qui peut limiter les objectifs d'apprentissage.

L'apprentissage par préférence propose de généraliser la construction des récompenses en utilisant un modèle probabiliste de Bradley-Terry [105, 10]. Ce modèle permet de convertir les préférences des médecins, basées sur la différenciation des états des patients, en récompenses quantitatives et ordinales. Dans la deuxième partie du Chapitre 4, nous présenterons notre méthode de construction des récompenses par apprentissage par préférence. Ce processus se déroule en trois étapes : (1) un expert exprime des préférences entre des paires d'éléments, ce qui induit un classement parmi toutes les instances du jeu de données collecté ; (2) les récompenses sont ensuite construites à l'aide du modèle probabiliste de Bradley-Terry ; (3) ces récompenses sont utilisées pour apprendre la politique dans les modèles de *Q-learning*. La principale contribution de cette méthode réside dans sa capacité à construire des récompenses de manière généralisée et guidée par les données. Cette approche exploite non seulement l'expertise des professionnels de santé, mais également les relations entre les données des patients, évitant ainsi les constructions manuelles de récompenses qui peuvent entraîner des choix arbitraires, tout en garantissant une cohérence dans l'apprentissage des stratégies médicales. Cette méthode sera illustrée à travers deux études de cas : l'une portant sur le traitement des adolescents atteints d'obésité [8, 61] et l'autre sur une simulation de cancer [144].

1.3 *Outcome-Weighted Learning*

1.3.1 Une méthode de classification pondérée

Pour construire des règles de décision médicale, nous avons abordé une famille de méthodes centrées sur l'apprentissage par renforcement, avec l'algorithme de *Q-learning* au cœur des applications DTR. Comme mentionné précédemment, il existe également une deuxième famille de méthodes, plus directe et non paramétrique, parmi laquelle se trouve la méthode *Outcome-Weighted Learning* (OWL). Cette méthode a été plus particulièrement développée dans le cadre de décisions à une seule étape, les ITR, où l'on étudie la possibilité de choisir entre deux traitements, tels que $A \in \mathcal{A} = \{-1, 1\}$. Dans l'article [143] qui présente cette méthode, il a été montré que déterminer un ITR optimal équivaut à résoudre un problème de classification. Plus précisément, ils reformulent la recherche de stratégies médicales en un problème de classification pondérée, où la frontière de décision représente la règle de décision entre les deux traitements et les poids sont déterminés à partir des réponses des patients. D'un point de vue d'apprentissage automatique, il s'agit d'un problème de classification à deux classes où les labels correspondent aux traitements administrés, et les observations i pour chaque patient sont pondérées par la récompense observée R_i et la propension ρ à recevoir le traitement. Ce problème est ensuite résolu à l'aide d'algorithmes tels que les *support vector machines*.

1.3.2 Quantification d'incertitude et *Bayesian OWL*

La recherche en statistique pour la médecine de précision se structure autour de trois grands axes principaux. Le premier axe est l'estimation : à partir d'une stratégie d, pouvons-nous en estimer sa valeur spécifique ? Le deuxième axe est l'optimisation de d : peut-on identifier une stratégie optimale d^{opt} ? De nombreuses méthodes abordées précédemment s'efforcent de répondre à cette question, et des extensions continuent d'être développées pour améliorer encore davantage cette optimisation. Le troisième axe est l'amélioration du recueil de données : afin d'estimer efficacement la valeur d'un régime d et de déterminer d^{opt}, quelles données sont nécessaires ? De nombreuses études se concentrent sur le design et la construction d'essais cliniques pour répondre à la question : quel est le design optimal pour générer des données pertinentes dans le cadre des analyses de médecine de précision ?

Grâce aux avancées dans ces trois domaines et aux preuves d'amélioration des soins qu'elles apportent, le besoin d'intégrer ces méthodes de détermination de stratégies médicales personnalisées dans la pratique clinique quotidienne devient de plus en plus pressant. D'une part, cela passe par le souhait d'incorporer ces nouvelles méthodes dans le système de santé tout en démontrant leur efficacité [46], et d'autre part, par le développement d'outils statistiques exploitables et adaptés à une utilisation par les professionnels de santé. C'est dans cette seconde thématique que s'inscrit la méthode développée dans le Chapitre 5 : *Bayesian OWL*. En effet, la formulation de la méthode OWL dans un cadre bayésien permet d'utiliser les méthodes d'inférence bayésienne pour quantifier l'incertitude sur les traitements recommandés. En partant de la fonc-

tion objectif de l'OWL, nous générons une pseudo-vraisemblance qui peut être exprimée comme un mélange d'échelles de distributions normales. Un algorithme d'échantillonnage de Gibbs est développé pour échantillonner la distribution postérieure des paramètres. Une fois transformée d'un cadre d'optimisation à un cadre probabiliste, notre méthode génère une distribution postérieure complète, utilisable pour l'inférence et, plus important encore, pour la quantification de l'incertitude des recommandations de traitement. Cela constitue un outil précieux pour les professionnels de santé dans l'aide à la décision thérapeutique.

1.4 Organisation du manuscrit

Après les deux chapitres d'introduction, respectivement en français et en anglais, le Chapitre 3 s'ouvrira sur la présentation du cadre mathématique de l'apprentissage par renforcement appliqué à la médecine de précision. Nous y définirons des concepts fondamentaux tels que les processus décisionnels, les politiques, les récompenses et les fonctions de valeur. Cette section se conclura par une illustration du *Q-learning*, l'un des algorithmes les plus couramment utilisés en RL, à la fois dans son format classique d'apprentissage en ligne et dans son application aux données observationnelles sous une forme rétroactive. Nous aborderons ensuite le contexte multi-étapes des DTR, en mettant l'accent sur le lien entre le formalisme du RL et les propriétés des algorithmes de RL adaptés à ce cadre. Enfin, nous introduirons le contexte à une seule étape des ITR, en les positionnant par rapport aux DTR et au RL. En utilisant le formalisme simplifié des ITR, nous présenterons les concepts clés et les hypothèses relatives à la causalité en médecine de précision.

Dans le chapitre 4, nous commencerons par présenter un état de l'art des méthodes permettant d'améliorer l'apprentissage par renforcement dans le contexte médical en intégrant les connaissances d'experts. Diverses méthodes seront exposées, expliquant les techniques par lesquelles elles intègrent l'expertise médicale et quelles parties du modèle d'apprentissage par renforcement sont modifiées pour incorporer ce savoir. Nous prendrons également le temps d'identifier les propriétés de chacun de ces algorithmes afin de faire un parallèle avec les propriétés attendues et idéales dans le contexte des DTR. Dans la seconde partie de ce chapitre, nous présenterons notre méthode d'apprentissage des récompenses par apprentissage des préférences, conçue pour une application aux DTR. Cette méthode sera illustrée par deux exemples d'application : l'une sur le traitement des adolescents atteints d'obésité [8, 61] et l'autre sur une simulation de cancer [144]. Ce chapitre se conclut par une section mettant en lumière les perspectives de recherche sur l'intégration des connaissances médicales dans les modèles d'apprentissage par renforcement, et plus particulièrement sur la continuité des recherches alliant apprentissage par préférence et apprentissage par renforcement.

Dans le chapitre 5, nous introduisons une approche bayésienne de la méthode OWL. À notre connaissance, il s'agit de la première stratégie bayésienne visant à apprendre directement des ITR optimaux. Nous utilisons une construction similaire à celle de [88], en construisant une pseudo-vraisemblance à partir de la fonction de perte pondé-

rée de classification. Nos principales contributions dans ce chapitre sont les suivantes. Tout d'abord, nous proposons une approche bayésienne pour apprendre des ITR optimaux, en utilisant un cadre basé sur la classification. Ensuite, nous développons un algorithme simple d'échantilloneur de Gibbs pour apprendre ces ITR optimaux. Enfin, nous démontrons comment utiliser la distribution pseudo-postérieure obtenue pour quantifier l'incertitude dans les recommandations de traitement. Nous démontrons les performances de notre approche à travers des études de simulation. Nous conclurons ce chapitre en abordant les améliorations possibles et les perspectives de recherche pour notre méthode *Bayesian OWL*, en mettant en lumière les directions futures pour perfectionner et étendre cette approche.

Enfin, le Chapitre 6 nous permettra de tirer une conclusion globale de ce manuscrit en récapitulant les principales contributions et résultats de notre recherche. Nous discuterons des implications de nos travaux et la détermination de règles de décision optimales pour la médecine de précision, en mettant en évidence les avancées réalisées ainsi que les défis restants.

Chapter 2

Introduction

2.1 Precision Medicine

2.1.1 Overview

Medical research has significantly enhanced our understanding of the biological mechanisms that influence the progression and development of chronic diseases, mechanisms that vary considerably from one patient to another. Thanks to the remarkable advances in modern medicine in terms of care, medications, and treatments, the exploration of new approaches is increasingly focused on the idea of providing the right treatment to the right person at the right time. This is the paradigm of precision medicine, also known as personalized medicine, which aims to tailor treatments to the individual characteristics of patients [11, 50, 51]. The goal is not to replace existing treatments or to develop a unique drug for each patient, but to complement the current therapeutic arsenal to enable personalized medical decision-making, with the aim of treating each person effectively. This is made possible by technological advances in data collection and storage. The volume of individual data collected has increased significantly, allowing for a better understanding of the individual factors influencing the effects of an intervention. These new and extensive databases allow for the control of patient heterogeneity. Each individual is unique, whether in terms of genetics, environment, lifestyle, or particularly in their response to treatments. These variabilities, often highlighted by randomized controlled clinical trials, underscore the importance of implementing more personalized treatments, thereby improving the quality of care provided.

We focus here specifically on chronic diseases like cancer, diabetes, or psychiatric disorders... Patients with these conditions go through long-term treatments that are regularly checked or adjusted by doctors. Precision medicine employs decision rules to recommend treatments specifically based on the patient's current condition. These decision rules are formalized through Dynamic Treatment Regimes (DTR). A DTR is a sequence of decision rules, one for each step of treatment, that guides how to personalize treatments based on the patient's treatment history and other factors [11]. When there is only one step in the medical intervention, with just one decision point, it is called an Individualized Treatment Regime (ITR).

The search for personalized medical decision rules, especially within DTRs, relies on longitudinally structured observational medical data. Analyzing this observational data in medicine requires considering causality. In precision medicine, understanding causal relationships is crucial to determine how medical interventions individually affect patients. Unlike simple correlations, causal relationships help identify the direct

effects of treatments on health, allowing clinicians to tailor treatments to the specific characteristics of patients, ensuring appropriate and effective interventions. To draw causal conclusions from observed data, it is crucial to adhere to certain assumptions, which we will detail in Section 3.4.3 of Chapter 3. One clinical trial design that meets these conditions is the Sequential Multiple Assignment Randomized Trials (SMART). SMART trials are experimental studies where patients are randomized at multiple decision points to test and optimize adaptive treatment strategies. Considered the "gold standard" of clinical trials in precision medicine, these designs are widely studied in the literature [13, 51, 11]. However, they are complex and expensive to implement, limiting the number of available trials. Notable examples include CATIE [107], ADHD [11, 54], and STAR*D [11, 53].

2.1.2 Individualized treatment regime

A formal framework for decision-making in precision medicine is the concept of ITR. This approach considers a single decision point based on observed data of size $n \in \mathbb{N}$, in the form $\{(X_i, A_i, R_i)\}_{i=1}^n$, where $X \in \mathcal{X}$ represents the initial characteristics of the patient, $A \in \mathcal{A}$ is the administered treatment, and $R \in \mathbb{R}$ is the treatment response, with higher values indicating a better patient outcome. An ITR is a function $d : \mathcal{X} \to \mathcal{A}$. Under the regime d, patients with $X = x$ would be assigned the treatment $d(x)$ [50]. Identifying an optimal ITR involves finding the decision rule that maximizes the treatment response R for all patients.

2.1.3 Dynamic treatment regimes

The second formal framework, which generalizes ITR into a context with multiple decision points, are DTRs. The data is then represented as $\{(X_{1,i}, A_{1,i}, R_{1,i}, \ldots, X_{T,i}, A_{T,i}, R_{T,i})\}_{i=1}^n$, which consists of n identically and independently distributed replicates of $(X_1, A_1, R_1, \ldots, X_T, A_T, R_T)$, where $X_1 \in \mathcal{X}_1$ are the patient's characteristics at the beginning of the study, and $X_t \in \mathcal{X}_t$ are the data collected at the intermediate time t for $t = 2, \ldots, T$; $A_t \in \mathcal{A}_t$ is the treatment assigned at time t; R_t is the response measured at time t. We also introduce $H_1 = X_1$ and $H_t = (H_{t-1}, A_{t-1}, R_{t-1}, X_t)$ where H_t represents the patient's medical history at time t. A DTR is a sequence of functions $d = (d_1, \ldots, d_T)$ where $d_t : \mathcal{H}_t \to \mathcal{A}_t$, meaning that at each decision point, a treatment is recommended based on the treatment history. An optimal DTR maximizes the expected cumulative response measure of the vector $\mathbf{R} = (R_1, \ldots, R_T)$.

2.1.4 Decision rule construction methods

To construct these optimal decision rules, two main families of approaches exist. The first family relies on regression models such as Q-learning [126, 75] and G-estimation [96, 99, 95]. The second family includes methods that estimate the expected value of the response under a particular regime without making parametric assumptions. Once this value is estimated, the optimal decision rule can be determined by exploring a class of possible regimes. Among these methods are marginal structural

mean models [97, 83], inverse probability of treatment weighting [97], augmented inverse probability of treatment weighting [139], and Outcome-Weighted Learning (OWL) [143].

Our research relied on a method from each resolution family: Q-learning and OWL. In the remainder of this introduction, we will first focus particularly on Q-learning within the DTR framework, an algorithm that is part of the broader field of reinforcement learning. Indeed, we will see that this subdomain of machine learning, which addresses multi-stage decision problems, can be perfectly adapted to the context of precision medicine in Section 2.2. This will allow us to situate two of our research works: the integration of medical knowledge into reinforcement learning models and the construction of data-driven rewards through preference learning. Then, we will turn our attention more specifically to one of the non-parametric methods applied in the ITR framework: the OWL method in Section 2.3. This will lead us to introduce our third contribution, which offers a Bayesian perspective for OWL, enabling the quantification of uncertainty associated with the recommended treatments.

2.2 Reinforcement learning

2.2.1 What is reinforcement learning?

Q-learning, mentioned among the previous methods, is part of a broader set of techniques known as Reinforcement Learning (RL). This area of research in machine learning focuses on acquiring control strategies for a system by exploiting the dynamic interactions between this system and an intelligent agent. The goal is to enable this agent, a learning entity, to maximize a predefined objective through its interactions with the environment. Typically, RL takes place in an interactive environment. The agent selects actions that lead to changes in the state of the environment. In return, the environment provides rewards, specific numerical values that can be positive or negative. These rewards indicate how beneficial or harmful the agent's actions have been concerning the overall goal of maximizing cumulative rewards. Through this feedback, the agent adjusts its strategy to improve its future decisions and more effectively achieve its objective. Thus, the field of RL introduces the following concepts:

- **State**: a representation of the current situation in which the agent finds itself. States contain the necessary information to decide on future actions.
- **Action**: a decision made by the agent that affects the environment. Actions can be discrete (such as moving a piece on a board) or continuous (such as adjusting the speed of a car).
- **Reward**: a signal received by the agent after performing an action. The reward indicates how beneficial the action was concerning the agent's objective.
- **Policy**: a strategy or set of rules that the agent uses to choose its actions based on the states. The policy can be deterministic (a fixed action for each state) or stochastic (a probability distribution over actions).
- **Value functions**: an estimate of the expected future cumulative reward that an agent can expect from a given state or action. Value functions help evaluate

policies and determine the best actions.

The foundations and origins of RL are attributed to Richard Bellman in the 1960s [6, 7, 108] through his work on dynamic programming and optimality, whose equations now form one of the key principles of this field. In the 1990s, the Q-learning algorithm was proposed by Chris Watkins, marking a significant advancement [125]. Simultaneously, Richard Sutton developed the temporal difference learning method [112]. In 1998, Andrew Barto and Richard Sutton published the first edition of "Reinforcement Learning: An Introduction" [113], which summarizes and compiles their previous research and developments. This book, often cited as an essential source for understanding the evolution and foundations of reinforcement learning, lays the groundwork for the concepts of reward and policy formalized within the framework of Markov decision processes.

The classical mathematical formalism widely adopted by the reinforcement learning community is that of Markov Decision Processes (MDP). An MDP is defined by a quintuple (S, A, P, R, γ) [9, 28] where:

— **States**: the set of possible states of the system, S.
— **Actions**: the set of actions available to the agent, A.
— **Transition probabilities**: function $P(s'|s, a)$ giving the probability of reaching state s' after taking action a in state s.
— **Reward**: function $R(s, a)$ providing the reward received after taking action a in state s.
— **Discount factor**: parameter $0 \leq \gamma \leq 1$ that weights the importance of future rewards.

The objective is to find an optimal policy π^* that maximizes the expected long-term cumulative reward from each state s.

RL finds applications in various domains due to its ability to solve complex sequential decision-making problems. In games and simulations, it is used to train agents in complex games like Go and chess. In robotics, RL enables robots to navigate, avoid obstacles, and manipulate objects with precision. In industrial automation, it optimizes production lines and resource management in real-time. In finance, it aids in developing algorithmic trading strategies and optimizing risk management. In transportation and logistics, RL is employed to train autonomous vehicles and optimize supply chains. It also enhances content personalization in recommendation systems. In the context of this thesis, we are particularly interested in the application of RL in healthcare, which has sparked significant interest in medical research through numerous reviews [133, 16, 21]. RL can serve healthcare in various ways, from logistics and hospital management to patient care. Here, we will focus specifically on its use in precision medicine for learning personalized treatment strategies.

2.2.2 Reinforcement learning and precision medicine

Machine learning has seen significant development over the past few decades, offering effective solutions for solving complex, large-scale problems. Among these approaches, RL has stood out for its ability to learn decision rules in sequential scenarios,

demonstrating its utility in precision medicine. RL primarily aims to identify optimal decision rules by maximizing long-term cumulative gains. In medicine, where the effects of treatments and their potential side effects may manifest after several stages, developing a long-term strategy is a major advantage. Additionally, RL models allow for the simultaneous processing of large sets of patient covariates while addressing multi-stage decision-making problems.

The modeling of the underlying sequential decision-making problem establishes a link between DTR and RL. As a result, RL has become widely recognized as one of the leading methods for determining DTRs in recent years [127, 19, 121, 15, 60]. However, the transition from RL to DTR, while apparent, is not always straightforward. Classical RL, as it is typically presented, is not necessarily well-suited for addressing the specific challenges of medical applications, and some algorithms are better equipped to meet the demands of this field. First, describing a DTR through a Markov decision process can be limiting due to the Markov assumption. This assumption stipulates that the most recent patient state contains all the necessary information for decision-making, which requires disregarding the patient's medical history. This is a strong assumption in the healthcare setting, where it is often preferable to consider a more general decision-making process. Second, RL is generally introduced in the context of an interactive environment. However, it is both impossible and unethical to allow an algorithm to interact with patients in a trial-and-error manner to learn an optimal treatment strategy. Instead, this strategy must be determined from data that has already been collected. This situates us in a particular context of RL applied in an offline mode. Specifically, the data has been collected according to a pre-established medical strategy, either by a physician or dictated by a clinical trial. We then learn an optimal strategy relative to the one that was already followed, which is referred to as off-policy RL. RL is governed by various dichotomies: model-based/model-free, policy-based/value-based, and on-policy/off-policy. These properties can help us navigate the vast RL algorithmic literature. This brings me to my third point on the specificities of applying RL to DTR: Q-learning. Among the wide variety of algorithms, backward Q-learning has emerged as the ideal algorithm for applying to DTRs (see Appendix 6). The properties of Q-learning, studied in detail, demonstrate why this algorithm is particularly well-suited to the needs of precision medicine.

The link between DTRs and RL, as well as the points raised here, will be discussed in Chapter 3, allowing for a deeper exploration of the RL formalism, the specificities of the application context, and the algorithmic properties.

2.2.3 Integrating medical knowledge into reinforcement learning models

While RL presents a promising technical solution for precision medicine, it relies on machine learning techniques that may raise concerns among both patients and practitioners. Ensuring that the decision rules constructed are safe, interpretable, and medically effective is one of the major challenges of applying RL in a hospital setting [21, 133]. One way to address this challenge is to integrate medical knowledge into

RL models. By combining clinical knowledge with machine learning algorithms, it is possible to create more robust models that are better adapted to real-world conditions. The idea is to create a synergy between the learning capabilities of machines and the expertise of domain professionals [39, 69]. This collaboration enhances trust in RL models and their recommendations [66], while also facilitating the adoption of this technology by healthcare professionals and patients in clinical settings [40]. The combination of machine learning and human expertise yields superior results compared to the use of RL alone or decisions made solely by experts [5, 56]. Moreover, from a technical perspective, the integration of medical knowledge reduces learning time, allowing for faster adaptation and improvement of methods, leading to more effective, patient-centered healthcare solutions.

The integration of medical knowledge into RL algorithms begins with the preparation of the model, particularly through data processing, data preparation, and the selection of the most appropriate algorithm. This integration specifically impacts key RL elements such as rewards, value functions, objective functions, and policy. The first part of Chapter 4 will present a state-of-the-art review of methods for integrating medical knowledge into reinforcement learning models, with a focus on the algorithmic properties chosen by each of these methods. This will highlight the necessary adaptations for the application of DTRs on observational data.

2.2.4 Reward construction through preference learning

In the context of integrating expert knowledge, preference learning is a particularly promising method. The idea is to construct the rewards of an RL model based on preferences provided by an expert. Rewards play a crucial role in learning strategies, as they constitute the criterion to optimize. Therefore, their design must encapsulate the state of the system as accurately as possible to provide meaningful guidance during learning. Typically, rewards are defined by a system expert who evaluates them using a score. For example, in clinical trials aimed at reducing weight in individuals with obesity, the reward may be measured by the body mass index [61]. In intensive care, treatments can be evaluated based on survival or mortality rates [101]. Some rewards are more complex and combine several variables. For example, in a cancer simulation presented in [144], rewards are based on tumor size, treatment toxicity, patient well-being, and survival rates. In the event of death, an arbitrary score of -60 is typically assigned. Manually constructing a reward function can involve arbitrary or highly context-specific choices, which may limit the learning objectives.

Preference learning proposes to generalize the construction of rewards using a probabilistic Bradley-Terry model [105, 10]. This model allows for converting the preferences of physicians, based on distinctions between patient states, into quantitative and ordinal rewards. In the second part of Chapter 4, we will present our reward construction method using preference learning. This process involves three steps: (1) an expert expresses preferences between pairs of elements, which induces a ranking among all instances in the collected dataset; (2) rewards are then constructed using the Bradley-Terry probabilistic model; (3) these rewards are used to learn the policy in Q-learning

models. The main contribution of this method lies in its ability to construct rewards in a generalized, data-driven manner. This approach leverages not only the expertise of healthcare professionals but also the relationships between patient data, thereby avoiding manual reward constructions that may lead to arbitrary choices, while ensuring consistency in the learning of medical strategies. This method will be illustrated through two case studies: one focusing on the treatment of adolescents with obesity [8, 61] and the other on a cancer simulation [144].

2.3 Outcome-Weighted Learning

2.3.1 A weighted classification method

To construct medical decision rules, we discussed a family of methods centered around reinforcement learning, with the Q-learning algorithm at the core of DTR applications. As mentioned earlier, there is also a second family of methods, more direct and non-parametric, among which is the Outcome-Weighted Learning (OWL) method. This method has been particularly developed in the context of single-stage decisions, such as ITR, where the possibility of choosing between two treatments is studied, for example, $A \in \mathcal{A} = \{-1, 1\}$, which represents the action or treatment. In their work, [143] demonstrated that determining an optimal ITR is equivalent to solving a classification problem. More specifically, they reformulate the search for medical strategies into a weighted classification problem, where the decision boundary represents the decision rule between the two treatments and the weights are determined based on patient responses. From a machine learning perspective, this is a two classes classification problem where the labels correspond to the administered treatments, and the observations i for each patient are weighted by the observed reward R_i and the propensity ρ to receive the treatment. This problem is then solved using algorithms such as support vector machines.

2.3.2 Uncertainty quantification and Bayesian OWL

Statistical research in precision medicine is structured around three main areas of focus. The first area is estimation: given a strategy d, can we estimate its specific value? The second area is the optimization of d: can we identify an optimal strategy d^{opt}? Many of the methods discussed earlier aim to answer this question, and ongoing extensions are being developed to further enhance this optimization. The third area is improving data collection: in order to effectively estimate the value of a regime d and determine d^{opt}, what data is necessary? Numerous studies focus on designing and constructing clinical trials to answer the question: what is the optimal design for generating relevant data in the context of precision medicine analyses?

With the advancements in these three areas and the demonstrated improvements in patient care they bring, the push to integrate these personalized medical strategy determination methods into everyday clinical practice is becoming increasingly urgent. On the one hand, this involves incorporating these new methods into the healthcare

system while demonstrating their effectiveness [46], and on the other hand, it requires developing statistical tools that are practical and suitable for use by healthcare professionals. It is within this second focus area that the method developed in Chapter 5, Bayesian OWL, is situated. By formulating the OWL method within a Bayesian framework, we can leverage Bayesian inference methods to quantify the uncertainty of the recommended treatments. Starting from the OWL objective function, we generate a pseudo-likelihood that can be expressed as a scale mixture of normal distributions. A Gibbs sampling algorithm is developed to sample the posterior distribution of the parameters. Once transformed from an optimization framework to a probabilistic one, our method generates a complete posterior distribution, which can be used for inference and, more importantly, for quantifying the uncertainty of treatment recommendations. This constitutes a valuable tool for healthcare professionals in therapeutic decision-making.

2.4 Structure of the manuscript

Following our two introductory chapters, presented in French and English respectively, Chapter 3 will introduce the mathematical framework of reinforcement learning as applied to precision medicine. We will cover key concepts like decision processes, policies, rewards, and value functions. This section will end with an example of Q-learning, a widely used RL algorithm, shown both in its traditional online form and in its use with observational data in a retrospective manner. We will then look at the multi-step context of DTRs, focusing on how RL formalism connects with the properties of RL algorithms in this area. Finally, we will introduce the single-step context of ITR, comparing them to DTRs and RL. Using the simpler ITR framework, we will explain the key ideas and assumptions related to causality in precision medicine.

In Chapter 4, we will begin by presenting a state-of-the-art review of methods that enhance reinforcement learning in the medical context by integrating expert knowledge. Various methods will be discussed, explaining the techniques by which they incorporate medical expertise and how different components of the RL model are modified to include this knowledge. We will also take the time to identify the properties of each algorithm to draw parallels with the expected and ideal properties in the context of DTRs. In the second part of this chapter, we will present our preference-based reward learning method, designed for application to DTRs. This method will be illustrated through two case studies: one on the treatment of adolescents with obesity [8, 61] and the other on a cancer simulation [144]. Chapter 4 will conclude with a section highlighting future research perspectives on integrating medical knowledge into reinforcement learning models, with a particular focus on the continued exploration of combining preference learning with reinforcement learning.

In Chapter 5, we introduce a Bayesian approach to the OWL method. To our knowledge, this is the first Bayesian strategy aimed at directly learning optimal ITR. We use a construction similar to that of [88], building a pseudo-likelihood from the weighted classification loss function. Our main contributions in this chapter are as follows. First,

we propose a Bayesian approach to learning optimal ITR, using a classification-based framework. Next, we develop a simple Gibbs sampling algorithm to learn these optimal ITR. Finally, we demonstrate how to use the resulting pseudo-posterior distribution to quantify uncertainty in treatment recommendations. We showcase the performance of our approach through simulation studies. We will conclude this chapter by discussing possible improvements and research perspectives for our Bayesian OWL method, highlighting future directions for refining and extending this approach.

Finally, Chapter 6 will allow us to draw a comprehensive conclusion of this manuscript by summarizing the main contributions and results of our research. We will discuss the implications of our work and the determination of optimal decision rules for precision medicine, highlighting the advancements made as well as the remaining challenges.

Chapter 3

Reinforcement learning for dynamic treatment regimes

Contents

3.1 Introduction 21
3.2 Theoretical foundations of reinforcement learning 23
3.2.1 Decision process 23
3.2.2 Policy 25
3.2.3 Rewards, valuation and optimization of policies 26
3.2.4 Reinforcement learning 30
3.3 The multi-decision setting : dynamic treatment regimes 33
3.3.1 Dynamic treatment regimes 33
3.3.2 Decision process and dynamic treatment regimes 33
3.3.3 Specificities of the medical context 34
3.3.4 Real data application 35
3.3.5 Properties of RL applied to DTR 36
3.4 The single-decision setting : individualized treatment regime . 40
3.4.1 Individualized treatment regime 40
3.4.2 Decision process and individualized treatment regime 41
3.4.3 Causality 42
3.5 Conclusion 44

3.1 Introduction

Modern medicine, with its remarkable advancements in care, drugs, and treatments, now seeks to enhance its ability to deliver personalized treatments for each individual patient. The paradigm of precision medicine [50] initiates a profound consideration of this question. Precision medicine aims to optimize the quality of healthcare by tailoring the medical approach to match the specific and continually changing health condition of every individual patient. The heterogeneity among patients' populations and sub-populations leads to distinct reactions and, consequently, necessitates different treatment approaches. Initially, this research domain introduced statistical models [11, 50, 51] aimed at facilitating decision-making support. Naturally, with the advent of data storage and the computational power, machine learning methods [16, 133] have also begun to be applied to address this issue.

In this context, one of the growing interests of modern medicine is to adapt prescribed treatments to the individual data, unique characteristics and particular medical history of the patient. Precision medicine seeks to put the patient's own information at the center in order to improve their health. The motto behind is "The right treatment for the right patient (at the right time)". In a 2015 State of the Union address, President Obama announced a Precision Medicine Initiative to revolutionize how we improve health, research, and treat disease. The initiative defines precision medicine as "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person" [115]. In technical terms, Adaptive Treatment Strategies (ATS) or Dynamic Treatment Regimes (DTR) formalize the objective of enhancing the care pathway for patients by proposing an optimal and personalized treatment sequence. They aim to establish a decision rule at each stage of the care process. It conditions the treatment based on responses to previous prescriptions and medical history [11, 54]. The goal is to optimize the patient's long-term positive response to the sequence of treatment decisions while tailoring the treatment to their own medical information [51].

In the past decades, machine learning has emerged as a solution to large-scale and high-complexity problems. When it comes to decision support, particularly in sequential scenarios, Reinforcement Learning (RL) [113] offers the most effective solution. These methods excel in adapting to changing conditions and optimizing decisions over a series of steps, making them especially valuable in dynamic decision-making processes. The concept revolves around identifying a decision rule, referred to as policy, which is designed to optimize a long-term objective. This policy is crafted in order to make decisions over time that lead to the greatest cumulative benefit or outcome.

RL methods is thus an appealing candidate for precision medicine and has been intensely studied as a potential tool to guide medical decisions towards personalized medicine. First, the application of these methods to DTR is facilitated by modeling the underlying decision problem using a so-called Decision Process (DP), as detailed in Section 3.2. It is straightforward to express and establish connections between medical elements and its mathematical components. Second, the primary aim of RL is to identify this decision rule. In this context, there is a desire to establish this rule while maximizing long-term cumulative gains. In medicine, the effects of treatments and side effects are not immediate but can take several stages to manifest. The way the policy is constructed is a significant asset for precision medicine. Third, RL models have the capacity to simultaneously consider the extensive patient covariates data and address multi-stage decision problems. The scope of RL applications in precision medicine is in recent thematic reviews of major interest : a non-technical survey offering illustrations of RL applications in public health is proposed in [127]. More specifically, RL applications in the context of mobile health are presented in [19]. Two more technical reviews describe the methods for determining medical decision rules using off-policy RL approach [121], or more specifically with the use of Q-learning [15] and their empirical comparison with other estimation methods [60].

In this chapter, we will first outline the mathematical framework of RL as applied to precision medicine. We will define key concepts such as decision processes, policies,

rewards, and value functions. This section will conclude with an illustrative example of RL through one of the most commonly used algorithms, Q-learning, presented both in its classical online learning format and its application to observational data in a backward form. In the following section, we will delve into the multi-stage context of DTRs, particularly focusing on bridging RL formalism with the properties of RL algorithms in this setting. Finally, we will introduce the single-stage context of Individualized Treatment Regime (ITR), positioning it in relation to DTRs and RL. Using the simpler formalism of ITR, we will present fundamental concepts and assumptions regarding causality in precision medicine.

3.2 Theoretical foundations of reinforcement learning

This section aims to outline the mathematical framework of RL applied in the DTR field. Typically, RL is explained in the context of a Markov Decision Process (MDP) and its evolution into a Partially Observable Markov Decision Process (POMPD). However, in this context, a return is made to a decision-making framework without the inclusion of Markov assumptions, which is referred to as a decision process. Subsequently, fundamental concepts are introduced : policy, value function, and the notion of optimality.

3.2.1 Decision process

3.2.1.1 General statement

The modeling context revolves around the realm of decision-making. A foundation proposed is DP, which acts as the initial framework for DTR. It represents a dynamic system which evolves through time $t \in \mathbb{T}$. This system navigates within the space of states $\mathbb{S}$ by executing actions within the realm of possibilities defined by the space of actions $\mathbb{A}$. The collection of non-empty measurable subsets of $\mathbb{A}$, denoted as $\{\mathbb{A}(s)|s \in \mathbb{S}\}$, represents the feasible actions that can be undertaken when the system finds itself in a specific state $s \in \mathbb{S}$.

Definition 3.2.1 (Decision process)**.** A decision process $(S, A, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$ on $\mathbb{T}$ includes:

- A family S of $\mathbb{S}$-valued vectors of random variables $\{S_t, t \in \mathbb{T}\}$, $\mathbb{S}$ is called space of states.
- a family A of $\mathbb{A}$-valued random variables $\{A_t, t \in \mathbb{T}\}$, $\mathbb{A}$ is called space of actions.
- a family $\{\mathbb{A}(s)|s \in \mathbb{S}\}$ of non empty measurable subsets of $\mathbb{A}$, the set of realizable actions when the system is in the state $s \in \mathbb{S}$. The requirement is for $\mathbb{K} = \{(s, a)|s \in \mathbb{S}, a \in \mathbb{A}(s)\}$ to be a measurable subset of $\mathbb{S} \times \mathbb{A}$.
- a distribution ν on $\mathbb{S}$.

Remark 3.2.1. DP is initially characterized for Borel spaces $\mathbb{S}$ and $\mathbb{A}$. However, in most practical applications, these spaces typically have finite dimensions, context considered for the rest of the article.

Remark 3.2.2. S_t represents the state at time t. This is a vector that includes several covariates observed at this time.

Remark 3.2.3. In full generalities, $\mathbb{T}$ will be taken as continuous or discrete but for a sake of readability $\mathbb{T}$ will be a discrete space denoted by $\mathbb{T} = \{0 = t_0, t_1, \ldots, t_n, \ldots, \tau\}$, with τ representing either a finite ($\tau = t_N < \infty$) or infinite ($\tau = \infty$) value. For the sake of simplicity, the variables X_{t_n} will be indicated as X_n and X_τ as X_∞ in infinite horizon setting.

Definition 3.2.2. For any $n \in \mathbb{N}$, an admissible history at time n is a vector which contains the states traveled by the system together with the actions taken up to time n. The set of admissible histories at time n is denoted:

$$\mathbb{H}_0 = \mathbb{S} \qquad \mathbb{H}_n = \mathbb{K}^{n-1} \times \mathbb{S}$$

An element $h_n \in \mathbb{H}_n$ writes $(s_0, a_0, \ldots, s_{n-1}, a_{n-1}, s_n)$ where for all $0 \leq j \leq n-1$, $(s_j, a_j) \in \mathbb{K}$.

The point of main importance to deal with the decision process is to exhibit the probability to reach state s_{n+1} at time $n+1$ given the history up to time n and the decision taken at time n this expresses as:

$$\mathbb{P}_\nu \left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n \right]. \tag{3.1}$$

In practice the computation of these probabilities requires significant computational resources because of the increasing length of the vector h_n as n increases. Rapidly working directly with such variable is intractable (usually when $n \geq 4$).

3.2.1.2 Markov decision process

To overpass this difficulty the Markov assumption is of particular interest. It consists in simplifying the dependence on the past by considering that all the necessary information for is contained in the current state.

Definition 3.2.3 (Markov decision process)**.** A Markov decision process on $\mathbb{T}$ is a decision process $(\mathbb{S}, \mathbb{A}, \{\mathbb{A}(s) | s \in \mathbb{S}\}, \nu)$ satisfying:

$$\mathbb{P}_\nu \left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n \right] = \mathbb{P}_\nu \left[S_{n+1} = s_{n+1} \mid S_n = s_n, A_n = a_n \right]. \tag{3.2}$$

A MDP is thus governed by a family of probability transitions

$$P_{a_n}(s_n, s_{n+1}) = \mathbb{P} \left[S_{n+1} = s_{n+1} \mid S_n = s_n, A_n = a_n \right].$$

which is the probability that action a_n in state s_n at time $t_n \in \mathbb{T}$ leads to state s_{n+1} at time t_{n+1}.

The most traditional RL framework is MDP [6, 28]. The majority of optimizing application complete their decision models with the memory-less Markov assumption.

Remark 3.2.4. Behind MDP modeling, there is a strong assumption that all the information necessary for the decisions observed. In reality, states space can be noisy or incomplete. To overpass this assumption, Partially Observable Markov Decision Process (POMDP) model introduced in [72] provides a relaxation to this assumption. POMDP can be seen as a generalization of MDP and is broadly based on the same framework. The major difference comes from the expression of the state space. POMDP consider a distinction between observed data and unobserved data, whereas DP and MDP are based exclusively on the data which have been directly observed. Mathematically, POMDP defines as an MDP except S which is a family of $\mathbb{S}^{obs} \times \mathbb{S}^{unobs}$-valued random variables $\{(S_n^{obs}, S_n^{unobs}), n \in \mathbb{N}\}$ where S^{obs} is observed and S^{unobs} is not.

3.2.2 Policy

The crucial concept in addressing dynamic programming is the notion of a policy, which is formalized as follows :

A policy is a sequence $\pi = (\pi_n)_{n\in\mathbb{N}}$ of conditional distributions from $\mathbb{A}$ given $\mathbb{H}_n$ defined, for any $\mathcal{A} \in \mathcal{B}(\mathbb{A})$ and all $h_n \in \mathbb{H}_n$, by:

$$\pi_n(\mathcal{A}, h_n) = \mathbb{P}\left[A_n \in \mathcal{A} \mid H_n = h_n\right],$$

satisfying for all $n \in \mathbb{N}$, all $h_n \in \mathbb{H}_n$:

$$\pi_n(\mathbb{A}(s_n), h_n) = 1,$$

and for all $n \in \mathbb{N}$, all $h_n \in \mathbb{H}_n$ and all $a_n \in \mathbb{A}(s_n)$

$$\pi_n(a_n, h_n) > 0.$$

Decision-making is selecting an option based on environmental information. A policy represents a plan that establishes a sequence of actions. This strategy can be tailored to align with a specified objective. As a result, the focus will be on deriving the strategy that optimizes this objective. A policy π_n is a strategy that suggests, for every possible states $s_n \in \mathbb{S}$, an action $a_n \in \mathbb{A}(s_n)$ taking to account the history $h_n \in \mathbb{H}_n$ of the system.

Theorem 3.2.1 ([38, 82])**.** Given a policy π and the initial distribution ν, there is a unique probability $\mathbb{P}_\nu^\pi$ such that, for all $\mathcal{B} \in \mathcal{B}(\mathbb{S})$, the Borel algebra of $\mathbb{S}$, and $\mathcal{A} \in \mathcal{B}(\mathbb{A})$, the Borel algebra of $\mathbb{A}$:

$$\mathbb{P}_\nu^\pi\left[S_0 \in \mathcal{B}\right] = \nu(\mathcal{B}),$$
$$\mathbb{P}_\nu^\pi\left[A_n \in \mathcal{A} \mid H_n = h_n\right] = \pi_n(\mathcal{A}, h_n)$$

In the following, $\mathbb{E}_\nu^\pi$ denotes the expectation associated with the probability $\mathbb{P}_\nu^\pi$ for an arbitrary policy π and an initial distribution ν.

The following result is of major practical importance and expresses the likelihood to observe a trajectory h_n by means of the DP.

Theorem 3.2.2. Given $(S, A, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$ a decision process on $\mathbb{T}$ and π a policy, we have for all $n \in \mathbb{N}^*$ and all $h_n \in \mathbb{H}_n$,

$$\mathbb{P}_\nu^\pi [H_n = h_n] = \prod_{j=1}^{n} \mathbb{P} [S_j = s_j \mid A_{j-1} = a_{j-1}, H_{j-1} = h_{j-1}] \pi(a_{j-1}, h_{j-1}) \nu(s_0)$$

Proof. First, we have, for any $j \in \mathbb{N}^*$,

$$\begin{aligned}
&\mathbb{P}_\nu^\pi [H_j = h_j] \\
&= \mathbb{P}_\nu^\pi [S_j = s_j, A_{j-1} = a_{j-1}, H_{j-1} = h_{j-1}] \\
&= \mathbb{P} [S_j = s_j \mid A_{j-1} = a_{j-1}, H_{j-1} = h_{j-1}] \\
&\qquad \times \mathbb{P}_\nu^\pi [A_{j-1} = a_{j-1} \mid H_{j-1} = h_{j-1}] \mathbb{P}_\nu^\pi [H_{j-1} = h_{j-1}] \\
&= \mathbb{P} [S_j = s_j \mid A_{j-1} = a_{j-1}, H_{j-1} = h_{j-1}] \pi(a_{j-1}, h_{j-1}) \mathbb{P}_\nu^\pi [H_{j-1} = h_{j-1}]
\end{aligned}$$

Now, by induction over n, it is easily shown that,

$$\begin{aligned}
&\mathbb{P}_\nu^\pi [H_n = h_n] \\
&= \prod_{j=1}^{n} \mathbb{P} [S_j = s_j \mid A_{j-1} = a_{j-1}, H_{j-1} = h_{j-1}] \pi(a_{j-1}, h_{j-1}) \mathbb{P}_\nu^\pi [S_0 = s_0] \\
&= \prod_{j=1}^{n} \mathbb{P} [S_j = s_j \mid A_{j-1} = a_{j-1}, H_{j-1} = h_{j-1}] \pi(a_{j-1}, h_{j-1}) \nu(s_0).
\end{aligned}$$

□

In the framework of MDP, to follow the same lines as in the proof of Theorem 3.2.2, an additional assumption on the policy is needed yielding to the concept of Markov policy:

Definition 3.2.4 (Markovian policy)**.** [82] A Markovian policy $\pi = (\pi_n)_{n \in \mathbb{N}}$ is a policy satisfying for all $n \in \mathbb{N}$, all $\mathcal{A} \in \mathcal{B}(\mathbb{A})$ and all $h_n \in \mathbb{H}_n$:

$$\mathbb{P} [A_n \in \mathcal{A} \mid H_n = h_n] = \mathbb{P} [A_n \in \mathcal{A} \mid S_n = s_n] = \pi_n(\mathcal{A}, s_n).$$

3.2.3 Rewards, valuation and optimization of policies

3.2.3.1 Rewards

As discussed in the Introduction, the aim of DP modeling is to find optimal policies associated to an objective. To do so, a criterion of optimality has to be introduced. This criterion is usually built by means of rewards functions which provides a temporal judgment of the desirability of a state-action pair and are formalized as follows:

Definition 3.2.5. Reward is defined as a family of bounded $\mathbb{R}$-valued random variables $\{R_n, n \in \mathbb{N}\}$. For a sake of simplicity, let us denote for a given $n \in \mathbb{N}$, for all $h_n \in \mathbb{H}_n$, all $a_n \in \mathbb{A}$ and all $s_{n+1} \in \mathbb{S}$:

$$\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) = \mathbb{E}_\nu^\pi [R_{n+1} \mid H_n = h_n, A_n = a_n, S_{n+1} = s_{n+1}].$$

Remark 3.2.5. The concept of rewards functions are usually integrated in the definition of a decision process.

3.2.3.2 Valuation of policies and value-functions

State-value functions and state-action values functions are respectively known as V-function and Q-functions. These two concepts provide quantitative measures for evaluating policies, making meaningful policies comparisons and defining the optimal policy. These value-functions serve as qualitative evaluations for guiding strategic adaptations.

State-value functions allow to answer to : "How good is to be in state s after following the policy π?" while action-value functions allow to answer to : "How good it is to have done the action a following policy π knowing that they were in state s?". The key point is the evaluation is not assessing step-by-step evaluation but by means of the cumulative reward over time. In such a way, value functions focus on a long-term objective.

Definition 3.2.6. Given $\gamma < 1$ a discount parameter, the stage n long term discounted reward function is defined for all $n \in \mathbb{N}$, by:

$$G_n = \sum_{j=n+1}^{\infty} \gamma^{j-n-1} R_j$$

Definition 3.2.7 (Value functions [11, 103])**.** Given $(S, A, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$ a decision process on $\mathbb{T}$, $\{R_n, n \in \mathbb{N}\}$ a family of rewards, π a policy and $\gamma < 1$ a discount parameter.

— The stage n state-value function (V-function) for a history h_n is the total expected future rewards from stage n given by:

$$V_n^{\pi}(h_n) = \mathbb{E}_{\nu}^{\pi}\left[G_n \mid H_n = h_n\right].$$

— The stage n action-value function (Q-function) is the total expected future rewards starting from a history h_n, taking action a_n is given by

$$Q_n^{\pi}(h_n, a_n) = \mathbb{E}_{\nu}^{\pi}\left[G_n \mid H_n = h_n, A_n = a_n\right].$$

The crucial aspect to observe in these definitions is that, instead of a step-by-step evaluation, the approach aims to assess a long-term objective. The goal is to evaluate the cumulative reward over time. Asa consequence of a decision, after each time step t_n, an immediate reward R_n is received which is the most distinctive feature of RL. The value functions represent the total expected future reward starting at a particular state s_0 and thereafter choosing actions according to the policy π.

Remark 3.2.6. The discount factor γ introduced in the definition of the long-term reward at each step n aims to strike a thoughtful balance between immediate rewards and long-term rewards. It allows for a balancing between striving for the highest cumulative reward and the aim to reach substantial benefits within a reasonable time [16]. This is also a mathematical trick to make the sum converge.

Remark 3.2.7. In the finite horizon case $\tau = t_N$, the values functions can be defined in a similar way by considering

$$G_n = \sum_{j=n+1}^{N} \gamma^{j-n-1} R_j$$

Notice that in this framework, the introduction of a discount parameter is not needed and is usually fixed to 1 from the definitions.

Remark 3.2.8. To consider valuation in infinite horizon, we have considered processes in infinite horizon and to do so, the Markov assumptions on the decision process and on the policy are necessary. The discount factor is now mandatory to insure the convergence of the long term discounted reward. The values functions can be defined in the same way by considering conditional to S expectations:

$$V_n^\pi(s_n) = \mathbb{E}_\nu^\pi \left[G_n \mid S_n = s_n \right].$$
$$Q_n^\pi(s_n, a_n) = \mathbb{E}_\nu^\pi \left[G_n \mid S_n = s_n, A_n = a_n \right].$$

The following proposition highlights the link between V-functions and Q-functions.

Proposition 3.2.1 ([51, 103, 113])**.** For all $n \in \mathbb{N}$, all $h_n \in \mathbb{H}_n$ and $a_n \in \mathbb{A}$, we have:

$$V_n^\pi(h_n) = \sum_{a_n \in \mathbb{A}(s_n)} Q_n^\pi(h_n, a_n)\ \pi_n(h_n, a_n) \tag{3.3}$$

$$\begin{aligned} Q_n^\pi(h_n, a_n) = & \sum_{s_{n+1} \in \mathbb{S}} \left(\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma V_{n+1}^\pi((h_n, a_n, s_{n+1}) \right) \\ & \times \mathbb{P}_\nu^\pi \left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n \right]. \end{aligned} \tag{3.4}$$

The remaining issue consists in the computation of the value functions. To do so, the result of major importance is the recursive form of the value functions which states that the value functions can be decomposed into immediate reward plus discounted value of successor state.

Theorem 3.2.3 (Recursive form for value functions [11, 142])**.** For all $n \in \mathbb{N}$, all $h_n \in \mathbb{H}_n$ and $a_n \in \mathbb{A}$, we have:

$$\begin{aligned} V_n^\pi(h_n) = & \sum_{s_{n+1} \in \mathbb{S}} \sum_{a_n \in \mathbb{A}(s_n)} \left(\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma\ V_{n+1}^\pi(h_{n+1}) \right) \\ & \times \mathbb{P} \left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n \right] \pi(a_n, h_n) \end{aligned} \tag{3.5}$$

$$\begin{aligned} Q_n^\pi(h_n, a_n) = & \sum_{s_{n+1} \in \mathbb{S}} \left(\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) \right. \\ & \left. + \gamma \sum_{a_{n+1} \in \mathbb{A}(s_{n+1})} Q_{n+1}^\pi(h_{n+1}, a_{n+1}) \pi(h_{n+1}, a_{n+1}) \right) \\ & \times \mathbb{P} \left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n \right] \end{aligned} \tag{3.6}$$

Equations (3.5) and (3.7) are known as Bellman's equation. A policy being fixed, the Bellman equation can be solved, therefore making it possible to determine the values of the value functions and thus the values of Q-function. Indeed, in the case where the number of steps is finite, the Bellman equation actually hides a linear system of N equations to N unknowns, where N is final finite number of steps considered. It can therefore be solved, once translated into a matrix equation, by a technique such as the Gaussian pivot.

3.2.3.3 Optimization of the policies

The key concern of the RL problem is to determine the optimal policy, denoted as π^*, which represents the optimal strategy for maximizing our long-term reward function. In other words, it is about finding the best way to make decisions in an environment to obtain the highest long-term rewards. The search for the optimal policy is based on the Bellman optimality principle developed bellow.

Definition 3.2.8. The optimal state-value functions (V_n^*) are defined for all $n \in \mathbb{N}$, all $h_n \in \mathbb{H}_n$ as the maximum value functions over all policies

$$V_n^*(h_n) = \max_{\pi} \ V_n^{\pi}(h_n)$$

The optimal action-value functions (Q_n^*) are defined for all $n \in \mathbb{N}$, all $h_n \in \mathbb{H}_n$ and $a_n \in \mathbb{A}$, as the maximum action-value functions over all policies

$$Q_n^*(h_n, a_n) = \max_{\pi} \ Q_n^{\pi}(h_n, a_n)$$

Definition 3.2.9. Consider the partial ordering over policies defined by:

$$\pi' \geq \pi \quad \text{if and only if, for all } n \in \mathbb{N}, \text{ all } h_n \in \mathbb{H}_n, \quad V_n^{\pi'}(h_n) \geq V_n^{\pi}(h_n).$$

This partial ordering allows to define optimal policy in the following way:

Proposition 3.2.2. There exists an optimal policy π^* that is better than or equal to all other policies, $\pi^* \geq \pi$ for all π.

Theorem 3.2.4. All optimal policies achieve the optimal value functions and the optimal action-value functions, for all $n \in \mathbb{N}$, all $h_n \in \mathbb{H}_n$ and $a_n \in \mathbb{A}$,

$$V_n^{\pi^*}(h_n) = V_n^*(h_n) \qquad \text{and} \qquad Q_n^{\pi^*}(h_n, a_n) = Q_n^*(h_n, a_n).$$

Theorem 3.2.5 (Bellman Optimality Equations for Q_n^*)**.** For all $n \in \mathbb{N}$, all $h_n \in \mathbb{H}_n$ and $a_n \in \mathbb{A}$, we have

$$\begin{aligned} Q_n^*(h_n, a_n) = \sum_{s_{n+1} \in \mathbb{S}} & \left(\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma \max_{a \in \mathbb{A}(s_{n+1})} Q_{n+1}^*(h_{n+1}, a) \right) \\ & \times \mathbb{P}\left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n \right] \end{aligned} \tag{3.7}$$

Proof.

$$\begin{aligned}
Q_n^*(h_n, a_n) &= \max_{\pi}(\mathbb{E}_{\nu}^{\pi}[G_{n+1}|H_n = h_n, A_n = a_n]) \\
&= \max_{\pi}(\sum_{s_{n+1}\in\mathbb{S}} \mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma \sum_{a_{n+1}\in\mathbb{A}} Q_{n+1}^{\pi}(h_{n+1}, a_{n+1})) \\
&\times \mathbb{P}\left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n\right] \\
&= \sum_{s_{n+1}\in\mathbb{S}} [\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma \max_{\pi} \sum_{a_{n+1}\in\mathbb{A}} Q_{n+1}^{\pi}(h_{n+1}, a_{n+1})] \\
&\times \mathbb{P}\left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n\right]
\end{aligned}$$

Now, consider $a_{n+1}^* \in \arg\max_a Q_{n+1}^*(h_{n+1}, a)$, we have :

$$\begin{aligned}
Q_n^*(h_n, a_n) &= \sum_{s_{n+1}\in\mathbb{S}} [\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma \max_{\pi} Q_{n+1}^{\pi}(h_{n+1}, a_{n+1}^*)] \\
&\times \mathbb{P}\left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n\right] \\
&= \sum_{s_{n+1}\in\mathbb{S}} [\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma Q_{n+1}^*(h_{n+1}, a_{n+1}^*)] \\
&\times \mathbb{P}\left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n\right] \\
&= \sum_{s_{n+1}\in\mathbb{S}} [\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma \max_{a} Q_{n+1}^*(h_{n+1}, a)] \\
&\times \mathbb{P}\left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n\right]
\end{aligned}$$

□

As a consequence of the Bellman Optimality Equation, we can claim that an optimal policy can be found by maximizing over $Q_n^{\pi^*}(s, a)$ for all $n \in \mathbb{N}$ and by considering the optimal policy defined as

$$\pi_n^*(s, a) = \begin{cases} 1 & \text{if } a \in \arg\max_{a\in\mathbb{A}(s)} Q_n^*(s, a) \\ 0 & otherwise \end{cases} \tag{3.8}$$

Note that this policy is deterministic.

3.2.4 Reinforcement learning

The mathematical foundations established in the previous sections serve as the basis for building algorithms to determine decision rules. In the field of RL, numerous algorithms aim to learn optimal policies. We have chosen to present two of these algorithms to illustrate a first distinction between online and offline application contexts. Furthermore, the second algorithm presented has been widely adopted to meet our application context. A discussion on the different RL algorithms suitable for our context will be the subject of Section 3.3.5.

3.2.4.1 Forward Q-learning

Q-learning, proposed in 1989 by Chris Watkins [113, 125], is one of the most famous and widely used algorithms in RL. It was historically developed in the so-called online context where the algorithm can dynamically interact with its application context. This is associated with the notion of "agent" which is an entity capable of interacting with the environment while receiving rewards. The concept of interaction is related to the exploitation-exploration dilemma. The agent must, through trial and error, choose between exploiting acquired knowledge to maximize immediate rewards or exploring new actions to discover better long-term strategies [113]. An excellent illustration of this problem is the ϵ-greedy strategy presented in the following definition:

Definition 3.2.10 (ϵ-greedy Policy)**.**

$$\pi_\epsilon(s) = \begin{cases} \text{random action from } \mathbb{A}(s) & \text{with probability } \epsilon \\ \arg\max_{a \in \mathbb{A}(s)} Q(s,a) & \text{with probability } 1-\epsilon \end{cases}$$

where $\epsilon \in [0,1]$ is an hyperparemeter called the exploration rate.

Q-learning relies on the recursive Bellman equations (3.2.3). The idea is to estimate value functions based on the differences between current and previous estimates, and then to derive an optimal strategy from Equation (3.8) of Bellman optimality.

Algorithm 1 Q-learning

Initialisation : $Q(s,a)$ arbitrarily, set learning rate α, discount factor γ, and exploration rate ϵ
for each history to build **do**
  Initialize state s
  while s has not reached the terminal stage **do**
    Choose action a using policy derived from Q (e.g., ϵ-greedy)
    Take action a, observe reward r and new state s'
    Update $Q(s,a)$ using the Q-learning update rule:
      $Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$
    $s \leftarrow s'$
  end while
end for
Output: The optimal decison rule is determined such as $\pi^*(s,a) = \arg\max_a Q(s,a)$

3.2.4.2 Backward Q-learning

When exploration of the environment is challenging, learning can be conducted using existing data, allowing decision rules to be derived from a non-interactive environment. This is referred to as offline or batch-RL. In this context, the algorithm does not interact with its environment; learning relies on estimating value functions from pre-existing databases. This offline Q-learning [22, 84] follows a backward approach illustrated in Figure 3.1.

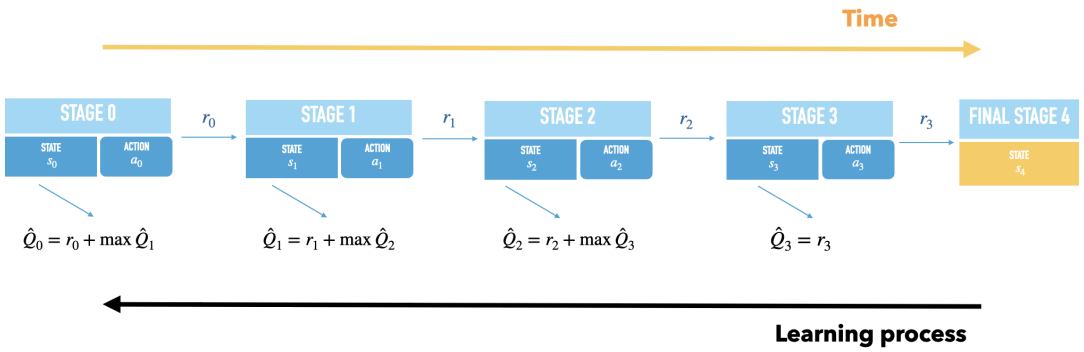

Figure 3.1 – Illustration of the Backward Q-learning algorithm for estimating Q-values on a history with 4 steps.

The estimates of the Q-function are initialized at the terminal time and move backward in time step by step. This strategy allows for the consideration of a possible delay effect commonly observed in longitudinal data. To estimate the Q-functions, various regression algorithms can be used, such as linear regression, support vector machines, decision trees or by deep neural networks, among others.

Algorithm 2 Backward Q-learning

Input: A set of training offline data consists of patients admissible histories h_t and their associated indexed reward r_t, $t = 0, ..., \tau$ and a regression algorithm
Initialisation : Let $t = \tau + 1$ and $\hat{Q}_t$ be a function equal to zero everywhere on $\mathbb{S} \times \mathbb{A}$
while until $t = 0$ **do**
$t \leftarrow t - 1$ (Backward)
Q_t is fitted with a regression algorithm though the following recursive equation : $Q_t(s_t, a_t) = r_t + \max_{a_{t+1}} \hat{Q}_{t+1}(s_{t+1}, a_{t+1})$
end while
Output: Given the sequential estimates of $\{\hat{Q}_0, ..., \hat{Q}_\tau\}$, the sequential optimal policies $\{\hat{\pi}_0, ..., \hat{\pi}_\tau\}$ can be determined

Remark 3.2.9. In an offline context, direct exploration is not present because decisions are made based on data collected in the database. Although there is no longer an exploration-exploitation dilemma as in the online context, it will be necessary to take into account a bias resulting from data where exploration-exploitation has already been performed.

3.3 The multi-decision setting : dynamic treatment regimes

3.3.1 Dynamic treatment regimes

Until the end of the 20th century, progress in medicine followed a "one-size-fits-all" approach. The search for the effect of a treatment or intervention was framed within evidence-based medicine on a target population. With the advent of massive data, particularly genomics, the paradigm has evolved. The volume of individual data collected has exploded, suggesting the possibility of integrating individual factors in the search for the effect of an intervention. The desired effect of treatment is no longer an average effect but a conditional effect on patient characteristics.

In this context, where the effect of an intervention is conditional to the variable characteristics of the patient which vary over time, the relevance of a treatment for a given individual may also vary over time. A central objective of precision medicine is to develop adaptive, and potentially optimal, intervention rules, where the definition of optimality must be clearly defined [47].

The search for adaptive (optimal) intervention rules is not a new question. A vast literature, primarily in the field of causal inference, exists and has real practical relevance. The foundational works in this context are attributed to [95], and the three extensions that allow for the effects of time-varying regimes in the presence of confounding variables: G-computation [93], the method of structural nested mean [94] models and G-estimation [96, 99, 95], as well as marginal structural models [97] and methods associated with inverse probability of treatment weighting [12]. Subsequently, a number of methods have been proposed, both in frequentist and Bayesian frameworks. All estimate the optimal DTR based on distributional assumptions of the data generation process via parametric models. We can consider them as direct resolution methods. These methods will not be further developed in this article; an up-to-date review including direct methods can be found in [18].

In the following section, we will detail the parallel that can be drawn between DTR and RL, which helps overcome a major barrier of direct methods, namely the risk of misspecification of underlying assumptions [142]. To address this limitation, in [77], followed immediately by [98], semi-parametric methods were considered, marking the first examples of RL-based approaches in the literature on DTR. The innovations of RL have breathed new life into the search for optimal DTRs, gradually expanding its applicability domain.

3.3.2 Decision process and dynamic treatment regimes

In Section 3.2, we notably introduced decision processes, policy and rewards which forms the theoretical foundation for algorithms searching for optimal policies, namely reinforcement learning. To describe the contribution of RL algorithms in the medical context, we will begin by examining how the framework introduced and DTRs are linked.

As discussed in Section 3.3.1, an adaptive intervention involves making a treatment decision based on the patient's characteristics and treatment history. An adaptive

decision rule can thus be perceived as a policy in the theoretical sense presented in Section 3.2.2. To leverage the results of reinforcement learning, it is essential to define the applied framework of the underlying DP for DTRs.

Building upon the definition 3.2.1 of a decision process, it is natural to consider, in a medical context:

— The state space $\mathbb{S}$ contains the selected covariates describing the patient's state.
— The action space $\mathbb{A}$ contains the selected treatments and their associated dosages.
— The subset $\{\mathbb{A}(s)|s \in \mathbb{S}\}$ states that the treatments feasible or accessible for a patient depend on a given state.

Remark 3.3.1. It is worth noting that in our context, the variable S_t is a vector containing a set of covariates observed at time t describing the patient's health state, which may influence the transition probabilities from one state to another.

The observed histories h_t are then the care pathways of different patients. They contain health data and treatments administered up to decision t.

One of the key elements of RL is the reward. In the medical context, rewards are defined to address the clinical objective. This is a very important point as optimization relies on it. The notion of reward will be central in the discussion on the integration of medical expertise in Section 4.2. Indeed, for a given situation, different rewards can be associated depending on the expertise of the physicians, the specific objectives of the clinical trial, either proximally (directly after the decision) or distally (at the end of the follow-up).

3.3.3 Specificities of the medical context

DTRs find their primary application in medical contexts where multiple treatment lines are possible or in contexts with multiple possible decision points (see Figure 3.2). These adaptive strategies are particularly relevant in areas such as intensive care, chronic diseases, psychiatry, or oncology.

Figure 3.2 – Illustration of medical history: treatment line for a patient.

The medical context is known for the great heterogeneity of its data [46, 111], whether in terms of care pathways, treatment response, side effects, social factors, or lifestyle. In this regard, data-driven methods offer interesting perspectives by overcoming the issue of model misspecification. Precision medicine would thus offer a path to more equitable access to treatments. Moreover, the decision-making process can take into account variables such as resource availability, finances, and other socio-economic or discriminatory factors, leading to fairer decision rules.

The timing of decision-making moments is a central issue in the problem of adaptive interventions. Typically, these decision points are linked to patient visits to the

practitioner. It is therefore natural to consider these moments as discrete and finite and to model them using a finite-horizon DP introduced in Section 3.2.1. Two issues arise: the time interval between two decision points and their frequency.

The issue of non-homogeneous time intervals between patients in the context of DTRs is typically addressed by considering the time between two visits as a covariate. Technically, this means defining the time based on the protocol and not worrying about the actual calendar time between visits. Even if the visits are not evenly spaced, by including this time information in a variable, we can treat the visits as if they are evenly spaced within the Markovian framework [54, 55, 103].

In some scenarios, such as patient follow-up in oncology or diabetes care, the number of visits is indefinite and varies based on individual patient needs. These patients are regularly monitored through mobile-Health (m-Health) initiatives, which operate in an online environment. Therefore, employing the Q-learning approach with backward induction, as explained in Section 3.2.4.2, becomes impractical. In [67], researchers identified optimal DTRs within an indefinite horizon framework using V-learning. This method aims to estimate the optimal policy from a predefined class of policies. Another approach, discussed in [23], utilizes an inferential procedure for estimating Q-functions.

Remark 3.3.2. In the rapidly expanding field of m-Health research, online approaches are particularly suitable. Just-In-Time Adaptive Interventions (JITAIs) have already been the subject of research efforts [43, 80, 92]. A synthesis of JITAI research is provided in [19], along with a comparative study with DTRs. This study addresses the technical aspect of making decisions about adaptive treatments in an interactive online environment. We will not cover these aspects further in the work, as the framework of DTRs on robservational data is discussed in Section 3.2.4.2, which is only feasible in the context of offline algorithms.

3.3.4 Real data application

Appendix 6 provides an overview of the RL research conducted in the context of DTRs. It is important to note that decision points are typically few in observational data application context; many studies consider two or three decision points. This choice is primarily driven by computational challenges: the more decision points there are, the more complex it becomes to integrate the patient's history into the models. An alternative approach is to impose a Markov assumption on the DP. However, in healthcare applications, this assumption is often unrealistic. The entire patient history can rarely be ignored or encapsulated in the current state.

As with any analysis on healthcare data, it is natural to question the biases inherent in the methods and the issue of causality [37, 81]. Since machine learning techniques are not causal inference methods, their use requires unbiased data. The issue typically arises in terms of "potential outcomes", and it is common to consider causal inference assumptions such as the "stable unit treatment value" assumption and the "no unmeasured confounders" assumption, as explained in [11, Chap. 2]. The question of causality in the field of reinforcement learning is also addressed more directly in the

framework of "causal RL" [1] [11, 140]. The search for adaptive intervention rules relies on data with a specific longitudinal structure. Innovations in algorithms for finding optimal DTRs often begin with adjustments to existing observational databases.

The Medical Information Mart for Intensive Care (MIMIC) [45] is a publicly accessible observational database containing information on 53,423 distinct admissions for patients in intensive care units between 2001 and 2012. It includes data on vital signs, medications, laboratory tests, measurements, caregiver notes, procedure and diagnostic codes, imaging reports, length of hospital stay, survival data, etc. Due to the wealth of available information and its longitudinal nature, MIMIC has been widely used by the RL community as a support for methods comparison (see [101], Table 3.1 and Appendix 6). It is also utilized as a training dataset for the development of data augmentation methods [118] and the generation of interactive environment models [86, 89].

Similarly to how randomized trials play a distinct role in clinical research and may be considered the gold standard for causal relationship investigation, the Sequential Multiple Assignment Randomized Trial (SMART) design [13, 51] can be regarded as the gold standard for clinical trial design in the context of adaptive interventions. SMART designs involve an initial randomization of patients to various treatment options, followed by re-randomizations at each subsequent stage of some or all of the patients to another available treatment at that stage. With such a design, the stable unit treatment value assumption is "by design" fulfilled. However, SMART designs are challenging to implement, costly, and as a result, there is limited access to data from SMARTs. However, notable trials include :

— CATIE (Clinical Antipsychotic Trials of Intervention Effectiveness) is a SMART study involving 1,460
schizophrenia patients over 18 months aimed at evaluating the clinical effectiveness of specific sequences of antipsychotic medications [107].

— ADHD (Attention Deficit Hyperactivity Disorder) is a SMART study involving 150 simulated participants, aimed at evaluating an adaptive intervention for children with this disorder. This study integrates behavior modification treatment along with medication treatment [11, 54].

— STAR*D (Sequenced Treatment Alternatives to Relieve Depression) is a SMART study involving 4,041 patients with major depressive disorders. This study evaluated the effectiveness of different treatment regimens [11, 53].

3.3.5 Properties of RL applied to DTR

There is a wide range of RL algorithms offering various methodological approaches tailored to specific contexts, as illustrated in Appendix 6 table. Figure 3.3 below provides a non-exhaustive overview of the most common RL algorithms. It presents many dichotomies, which will be explained in the following paragraph and contextualized in DTRs applications.

1. for details of "causal RL" initiative, see https://crl.causalai.net/

Reference	Model	State Space	Action Space	Rewards
[49]	SARSA	Discretised state space	25 unique actions based on a 5 by 5 binning procedure of maximum vasopressor dose and sum of intravenous fluids per 4h time interval	Terminal reward at the end of each trajectory based on 90-day mortality
[90]	Dueling DDQN	Ordinary and Sparse Auto-Encoders were used for latent state space representation	As paper [49]	Terminal reward at the end of each trajectory based on in-hospital mortality
[89]	Dueling DDQN	Continuous state space based on 4h aggregated features based on physiological parameters	As paper [49]	Intermediate reward based on changes in critical care scores and lactate combined with a terminal reward for survival based on ICU mortality
[86]	Dueling DDQN	Patient states are encoded recurrently using an LSTM autoencoder representing the cumulative history for each patient	As paper [49]	The change in the negative mortality logodds of mortality between the current observations and the next observations.
[58]	Actor-Critic	POMDP	As paper [49]	As paper [49]
[135]	Dueling DDQN	As paper [3]	As paper [49]	Developed several reward functions based on 7 potential features most important during the treatment process

Table 3.1 – Applications of RL algorithms on MIMIC database: highlighting various medical objectives with rewards design extract from [101].

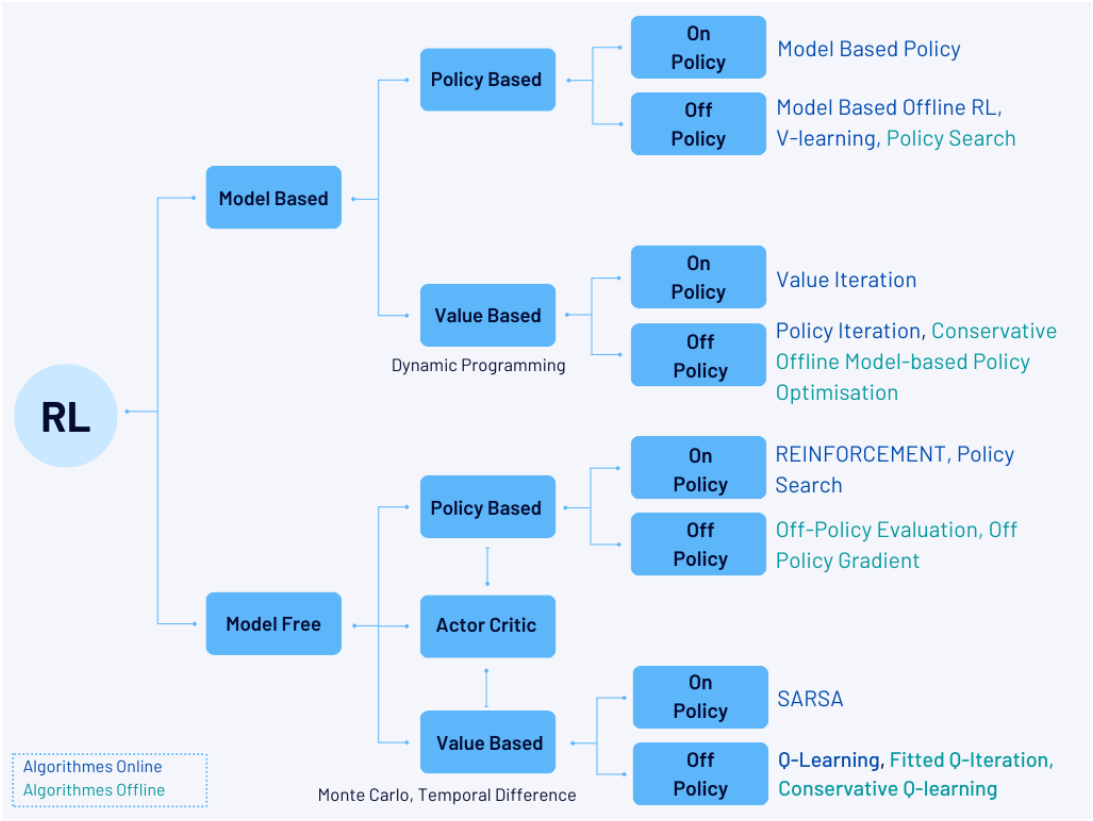

Figure 3.3 – Classification of the most common RL algorithms.

3.3.5.1 Model-based vs. Model-free

The first dichotomy in Figure 3.3 is based on the distinction between a model-based approach and a model-free approach. This distinction is related to the concept of transition probability defined by equation (3.1). A procedure is considered "model-based" when it relies on knowledge of all transition probabilities from a model, which means having access to all dynamics of the system. A model-free method is able to bypass this model and is based on partial information of the associations between states and actions to determine the optimal strategy. In a model-based approach, all possible paths from an initial state s_0 are explored, and an optimal policy is one that maximizes the objective.

However, in a medical context, exploring all possibilities from the same starting point is infeasible, mainly for clinical and ethical reasons. The environment is thus inherently partially observed. This reality inherently places us in a model-free framework. It is worth noting the existence of an application on simulated patient data based on MIMIC (see Section 3.1) in the model-based framework in [91].

3.3.5.2 Policy-based vs. Value-based vs. Actor-critic

The second distinction involves two different approaches to determine the best strategy: policy-based methods and value-based methods. The former aim to directly find the optimal policy by formalizing the RL problem through a family of policies, introduced in [113, Chapter 13]. The latter seek the optimal policy through value functions, introduced in Section 3.2.3.2, and serve as the basis for algorithmic methods such as dynamic programming, Monte Carlo, and temporal-difference, also presented in the same book. These two approaches can be combined, thus forming actor-critic methods [33, 113].

Policy-based Policy-based methods are direct approaches to finding the optimal policy that rely on a parametric form of the strategy π_θ for $\theta \in \Theta$. Optimization can be typically achieved through gradient descent :

$$\theta_{n+1} = \theta_n + \nabla \mathbb{E}_{\pi_\theta}[G_n|\theta] \tag{3.9}$$

where G_n is the cumulative long-term reward introduced in Remark 3.2.7.

This method has been applied to simulated HIV data [132] as well as in the intensive care domain [91]. Note that the first application highlighted the challenges of converging to an optimal decision rule due to the simplification of simulation models. The main obstacle to using this method is the difficulty of convergence, which requires a large volume of data.

Value-based Value-based methods evaluate the optimal policy indirectly based on value functions V^π or Q^π introduced in Section 3.2.3.2. The general idea is to quantitatively evaluate states or action-state pairs using one of the value functions (Q-function or V-function). The optimal policy is then obtained by identifying actions that maximize these values. The success of these methods relies on the ability to model these value functions, as outlined in Section 3.2.4.2, through algorithms such as Backward Q-learning, making it a highly flexible approach.

The initial work was conducted by [75], who introduced an offline Q-learning, also known as batch learning, in a context of non-Markovian planning with a limited and restricted number of steps ($n \leq 4$). This approach proves ideal for its application to DTRs and can serve as a starting point for many other applications. Research activity in this field quickly became significant, considering various parametric, semi-parametric, and non-parametric strategies to model the value function [11, 53, 76, 119].

Value-based methods are better suited for application to DTRs. They enable the discovery of optimal decision rules in a non-Markovian framework with a small number of steps and data, unlike policy-based methods. This makes them easily applicable to observational data. Moreover, they can offer a clearer interpretation, especially when Q-function estimation relies on a linear regression model [53], thus providing interpretable decision rules. As shown in Appendix 6, this is the most widely used method in practice, particularly Q-learning approaches and its derivatives in the context of DTRs.

Actor-critic A third approach to address the question of finding an optimal strategy is known as the 'Actor-Critic' method. It takes a hybrid approach by combining an actor based on policy-based methods with a critic based on model-based methods, thus integrating the advantages of both previous methods. The actor refines the parameterized policy under the guidance of the critic. The latter uses value functions, also parameterized V^{π_θ} or Q^{π_θ}, to guide learning. This third way of constructing decision rules was developed to correct biases in value-based methods and to counterbalance the high variability of the gradient part of policy-based methods in equation (3.9) [33].

Actor-critic methods have been applied to the MIMIC dataset. This compromise between policy-based and value-based methods converges towards a decision rule reducing patient mortality in [124] or providing a decision rule in line with physician's usual opinions in [57, 59]. This approach relies on gradient descent, similar to policy-based methods, thus necessitating databases containing a large number of individuals, often simulated data.

3.3.5.3 On-policy vs. Off-policy

This last dichotomy is closely related and sometimes confused with the concepts of offline and online algorithms presented in Section 3.2.4.

DTRs on observational data inherently operate in an offline context, aiming to determine the optimal policy using previously collected data. This means that, rather than adapting in real time, the analysis and optimization are done retrospectively. For example, all SMART designs gather data from established clinical protocols. These protocols dictate the timing and nature of patient visits, ensuring a structured collection of data. By analyzing this data, researchers can develop and refine treatment strategies [121, 11, 50]. Therefore, applying RL in the DTR context and clinical decision support is fundamentally off-policy, meaning that the strategy used to generate the data ('behavior policy') is not necessarily optimal. The optimal strategy ('target policy') is deduced subsequently.

On-policy algorithms require an interactive online context where the strategy generating the data is optimized. The concepts of behavior and target policies are merged. The online framework can benefit from both on-policy algorithms, as is the case in the medical domain with Just-in-Time Adaptive Interventions (JITAIs) discussed in [19], and off-policy algorithms (see Figure 3.3). Some online algorithms, both off-policy and on-policy, have been explored within the context of DTRs, but exclusively in simulated data settings, as indicated in Appendix 6.

3.4 The single-decision setting : individualized treatment regime

3.4.1 Individualized treatment regime

When introducing precision medicine in clinical decision-making, it is traditionally more common to start with a context involving a single medical decision step. This

is referred to as Individualized Treatment Regimes (ITR). However, the multi-stage context of DTR can be seen as a generalization of ITR. Indeed, ITR corresponds to a single decision or a single-stage problem. It is essentially a DTR with only one decision rule between two states [119].

This natural entry point into precision medicine is presented through a formalism of the form $\{(X_i, A_i, Y_i)\}_{i=1}^n$, where $X \in \mathcal{X}$ represents the baseline patient characteristics, $A \in \mathcal{A}$ is the administered treatment, and $Y \in \mathbb{R}$ is the outcome such that higher values indicate a better state of the patient. For each $x \in \mathcal{X}$, let $\psi(x) \subseteq \mathcal{A}$ represent the set of allowable treatments for a patient with $X = x$ [50]. More precisely, $X_i = (X_{i,1}, \ldots, X_{i,p})^T \in \mathcal{X}$ denotes the p-dimensional biomarker and prognostic information vector, a set of random variables contained in the space $\mathcal{X}$. We will consider the simple case of two treatments where the treatment space is binary $\mathcal{A} = \{0, 1\}$. This binary notation is associated with statistical resolution methods, but it is also common to find $\mathcal{A} = \{-1, 1\}$, whose reformulation benefits the formalism of other machine learning-based models.

An ITR in this context is a map $d : \mathcal{X} \to \mathcal{A}$ that satisfies $d(x) \in \psi(x)$ for all $x \in \mathcal{X}$. Under d, patients with $X = x$ would be assigned the treatment $d(x)$ [50]. Thus, fundamentally, identifying an optimal ITR involves finding the treatment rule that maximizes the expected outcome across all patients. The optimal regime is defined as d^{opt}.

As discussed in Section 3.3.1, the search for an optimal individualized medical decision dates back to the 1990s and initially focused on single-stage approaches. The two most well-known regression-based methods are Q-learning (see Section 1) and G-estimation [96, 99, 95]. Both methods primarily rely on the parametric estimation of $\mathbb{E}[Y|X, A]$. To move beyond linear decision rules and adopt a more direct and nonparametric approaches, the statistical research for precision medicine introduces methods such as marginal structural mean models [97, 83], inverse probability of treatment weighting [97], augmented inverse probability of treatment weighting [139], and outcome-weighted learning [143]. Each of these methods estimates the value of a specific treatment regime d, defined as $V^d = \mathbb{E}[Y(d)]$, which represents the expected average outcome if all patients were to follow regime d. Each method offers a distinct approach to weighting this value. A comprehensive explanation and comparison of these methods are provided in [51, Chapter 6].

3.4.2 Decision process and individualized treatment regime

In Section 3.2.1, we introduced fundamental concepts such as decision processes, policies, and rewards, which form the theoretical basis of RL. In Section 3.3.2, we established the connections between mathematical elements and their practical applications. In this section, we will perform a similar exercise, taking the time to draw parallels between the mathematical foundations of ITR and the RL framework.

Building on the definition of a decision process provided in 3.2.1, we can observe the following:

— An ITR represents a single-stage decision process, hence involving only a singu-

lar stage.
— The state space $\mathbb{S}$ encompasses the covariates that describe the patient. In the context of ITR, $\mathcal{X}$ and $\mathbb{S}$ coincide, meaning that X, the vector of baseline covariates, corresponds to S_0.
— The action space $\mathbb{A}$, which includes the available treatments, is represented by $\mathcal{A}$. This space is typically restricted to a limited number of treatment options, often just two.

Remark 3.4.1. In the context of a multistage discrete DTR problem, n refers to the number of stages. In contrast, in the ITR context, n denotes the number of patients.

In the context of ITR, the observed history h_t no longer describes longitudinal data. Instead, it consists solely of the covariates at the beginning of the trial and the administered treatment, represented as $h = (s_0, a_0) = (x, a)$.

The decision rule described in Section 3.4.1 can be generalized using the framework outlined in Section 3.2.2.

The final outcome Y serves as a quantitative measure of the patient's condition and thus corresponds to the reward. In some applications of ITR, Y may also be denoted as R, representing either the reward or the response.

Bridging the gap between the statistical framework of ITR and RL solving methods, such as Q-learning and other regression-based approaches, was addressed in [15].

3.4.3 Causality

The single-stage context with a binary action space provides an opportunity to discuss the causal framework using a simpler formalism, offering the initial key steps and intuitions, which can be generalized to a multi-stage framework.

In order to discuss how to make inferences from observational data concerning ITR, we need to introduce the notion of potential outcomes or counterfactuals. This concept refers to the patient's response if a certain treatment were administered (or if a certain regimen were followed), possibly different from the one actually observed (hence, counter to fact). A potential outcome $Y^*(a)$ is the outcome the patient would experience if they were to receive the treatment option a. We aim to determine the best treatment $a \in \mathcal{A}$ for a patient, corresponding to the largest $Y^*(a)$ for that patient. Obviously, it is impossible to identify the best treatment option for an individual since all potential outcomes for a given patient are not observable. Therefore, this problem cannot be solved at an individual level. However, it is possible to identify population-level causal parameters or average causal effects under conditions of perfect compliance with randomization, or to estimate bounds on these effects under conditions of non-compliance [11]. In the absence of randomization, such as in observational studies or randomized trials with imperfect compliance, additional assumptions are necessary to estimate population-level effects. These assumptions are presented in the single-stage setting, but applications to the multiple-stage setting can be found in [11, Chapter 2].

The axiom of consistency is a fundamental requirement stating that the potential outcome under the observed treatment and the actual observed outcome must be the

same. In other words, the expectation of the observed outcome when treatment a is administered is equal to the expectation of counterfactual outcome.

Remark 3.4.2. This axiom is often plausible in studies of medical treatments, where it is straightforward to design how to manipulate the treatments administered to patients. Consequently, it is easily verified by the study design and data collection process.

The three other assumptions are necessary for an unbiased estimation of treatment effects [11, 24] :

— The Stable Unit Treatment Value Assumption (SUTVA): A subject's outcome $Y(a)$ is not influenced by other subjects' treatment allocation [102].

— No Unmeasured Confounding (NUC): The covariates X encompass all information relevant for assigning treatments. This is expressed in the context by:

$$Y^*(a), Y^*(a') \perp A \mid X$$

This assumption implies that the effect of potential outcomes can be assessed using the data as follows:

$$\begin{aligned}\mathbb{E}[Y^*(a)] &= \mathbb{E}[\mathbb{E}[Y^*(a) \mid X]] \\ &= \mathbb{E}[\mathbb{E}[Y^*(a) \mid A = a, X]] \\ &= \mathbb{E}[\mathbb{E}[Y(a) \mid A = a, X]] \\ &= \mathbb{E}[Y(a)]\end{aligned}$$

— The positivity assumption, also known as Experimental Treatment Assignment (ETA), requires that every treatment option has a positive probability of being assigned given any set of covariates. This means that for each possible covariate-treatment pair, there must be a positive probability that the treatment prescribed by the treatment regime is observed. In other words, subjects should be able to receive both treatments without any restrictions.

Remark 3.4.3 (STUVA)**.** The SUTVA is often reasonable, particularly in the context of randomized trials. However, it may be violated in specific scenarios, such as vaccinations for contagious diseases, where "herd immunity" effects can influence the outcomes.

Remark 3.4.4 (NUC)**.** The NUC assumption is a standard but unverifiable requirement for observational studies. It is automatically satisfied for data obtained from randomized trials.

Remark 3.4.5 (Positivity)**.** In practice, the positivity assumption can be checked through studies of data distribution. However, positivity can be violated in two ways: theoretically or practically. A theoretical violation occurs when the study design prevents certain individuals from receiving specific treatments. A practical violation happens when certain groups of people have a very low chance of receiving the treatment [11, 24].

3.5 Conclusion

This chapter introduces and aims to facilitate the understanding of RL methods for precision medicine, especially its application to optimal DTR and ITR research. This topic is of major practical interest since it aims to determine an optimal decision rule for personalized treatments, with a large range of applications in areas such as intensive care, chronic diseases, psychiatry, and oncology. However, applying RL to medical research requires specific considerations and adaptations.

The main specificity arises from the data, typically derived from observational studies, which limits RL methods to offline applications. While an online setting is feasible, such as in m-health scenarios, for many cases, it is unethical to base treatment decisions solely on an algorithm. Therefore, since the data has already been collected beforehand, it is important to note that the well-known exploration-exploitation dilemma of online RL translates into an exploration-exploitation bias in offline RL settings. Section 3.3.5 details the properties of RL algorithms and helps identify the most desirable characteristics for an algorithm applied to DTR. First, due to clinical and ethical constraints, exploring all possibilities from the same starting point is impractical, necessitating the use of model-free algorithms. Secondly, value-based methods enable the discovery of optimal decision rules in a non-Markovian setting with limited steps and data, distinguishing them from policy-based approaches. Thirdly, off-policy algorithms are suited for offline contexts where data is already collected following a specific strategy, allowing for the determination of the optimal policy in a second phase. When these three characteristics converge, the result is an algorithm well suited for practical applications with observational DTR data. Consequently, Backward Q-learning, also known as Fitted Q-Iteration, emerges as the most widely adopted and utilized algorithm in the realm of applying RL to DTR [15].

Closely related to all research involving observational data, the issue of causality is also significant in the context of optimal DTRs. We have used the framework of ITR to provide initial insights and assumptions governing causality in this context, which can later be generalized to the domain of DTRs. However, when applying RL to DTRs, only a few studies address this challenge directly [11, 140], while most rely on assumptions that are difficult to verify, which can render the results questionable. This limitation can be mitigated by using experimental designs such as SMART, although these designs are complex and costly to implement [13, 51].

The classical formulation of RL relies on decision processes theory under the Markov assumption. However, this assumption is often too stringent in practical applications. Indeed, there is no guarantee that the current state under study contains all the necessary information to construct a precise decision. However most of the mathematical properties remain true without this Markov assumption by considering the entirety of the patient's history. In practice, that necessitates huge computational capacities and restricts to the applications the determination of adaptive strategies where the number of DTR steps is small (less than 4).

Chapter 4

Integrating medical knowledge into reinforcement learning models

Contents

4.1 Introduction 45
4.2 Approaches to integrating medical knowledge into RL 47
4.2.1 Medical knowledge and model preparation 47
4.2.2 Medical knowledge and rewards 48
4.2.3 Medical knowledge and value functions 49
4.2.4 Medical knowledge and objective function 51
4.2.5 Medical knowledge and policy 51
4.3 Rewards construction based on preference learning 52
4.3.1 Preference learning 53
4.3.2 BMI data application 55
4.3.3 Cancer application 61
4.3.4 Conclusion 66
4.4 Perspectives 68

4.1 Introduction

The previous chapter introduced Reinforcement Learning (RL) methods for precision medicine, particularly for optimal Dynamic Treatment Regimes (DTR) research. This is crucial for personalized treatments across various fields like intensive care, chronic diseases, psychiatry, and oncology. Applying RL in medical research requires specific considerations and adaptations. We noted that data typically comes from observational studies, limiting RL methods to offline applications. Note that, the exploration-exploitation dilemma of online RL becomes an exploration-exploitation bias. Online reinforcement learning can work in settings like mobile-health, where patients are frequently monitored. However, it is often unethical to rely only on algorithms for treatment decisions. Section 3.3.5 outlines the properties of RL algorithms, identifying key characteristics for DTR applications:

— **Model-Free:** due to clinical and ethical constraints, exploring all possibilities from the same starting point is impractical, making model-free algorithms essential.

— **Value-Based:** these methods help find optimal decision rules in non-Markovian settings with limited steps and data, unlike policy-based approaches.
— **Off-Policy:** suitable for offline contexts where data is collected following a specific strategy, allowing the determination of the optimal policy later.

When these characteristics align, the algorithm is well suited for practical DTR applications. Thus, backward Q-learning, or fitted Q-Iteration, is the most widely used algorithm for applying RL to DTR.

While RL offers promising algorithms for sequential decision-making in healthcare, as detailed in Appendix 6, relying on a machine learning algorithm may create apprehension among all stakeholders in the process. This hesitation can originate from both the patient and the physician sides. In order to be operational in a clinical context, several points must be improved such as safer, more interpretative and efficient medical decision making [21]. One approach to enhance the application of RL in healthcare is the integration of expertise or human knowledge into the models. The concept is to create a partnership between both machine learning capabilities and domain experts [39, 69]. This "collaboration" would not only improve confidence in RL models and the recommendations they provide [66] but also facilitate the utilization of this technology by healthcare professionals and patients within a clinical setting [40]. This merging of machine learning and human expertise yields to improved results compared to RL in isolation or expert decisions alone [5, 56]. From a technical point, involving experts or medical knowledge also reduces the learning time, allowing for quicker adaptation and enhancement of the methods, ultimately leading to more effective and patient-centered healthcare solutions.

The first objective of this chapter is to review the state of the art in integrating medical knowledge into reinforcement learning models. This involves paying particular attention to the algorithmic properties with which these models have been developed, thereby highlighting the necessary adaptations for their application to DTR derived from observational data.

One of the methods presented in subsection 4.2.2, and explored in depth, is the construction of rewards through preference learning [27]. Indeed, rewards are crucial elements in learning optimal strategies, as the overarching goal of RL is to maximize them. They provide quantitative indications of the system's state. Their construction or formulation is thus critical in decision-making learning. Generally, rewards are designed by a system expert who proposes to evaluate it through a score. For example, in the context of clinical trials for individuals with obesity, where the goal is to reduce their weight, the reward can be measured through their body mass index [61]. Another example, in critical care settings, is to evaluate treatments based on survival or mortality rates [101]. Some rewards can be designed more subtly by making compromises and combining variables. In the context of a cancer simulation presented in [144], rewards are evaluated based on tumor size, treatment toxicity, patient well-being, and survival rates. However, when a patient's death occurs, an arbitrary choice is usually made to assign a score of -60 to the event. Manually constructing a reward function can involve arbitrary or very context-specific choices and lead to overly restrictive learning objectives.

Preference learning offers an interesting approach to generalizing reward construction by employing a probabilistic Bradley-Terry model [105, 10] to convert physician preferences between patient trajectories into quantitative and ordinal rewards. The second objective of this chapter is to show how this method meets the ideal properties for applying RL to DTR: offline, model-free, value-based, and off-policy, while ensuring that learned policies align with medical objectives. This will be demonstrated through two case studies: one on treating adolescents with obesity [8, 61] and another on a generic cancer simulation [144].

To achieve the objectives of this chapter, we structure it as follows. In Section 4.2, we provide an overview of methods to enhance reinforcement learning in the medical context by integrating expert knowledge. Various methods are presented and discussed. In Section 4.3, we present our method of learning rewards through preference learning, designed for application to DTR, along with two examples of application. The chapter concludes with a section highlighting research perspectives in Section 4.4.

4.2 Approaches to integrating medical knowledge into RL

In Chapter 3, we examined the theoretical foundations of RL in Section 3.2. We discussed key concepts such as reward elements, value functions, and objective functions, which are essential for integrating medical knowledge into RL models. In this section, we provide a state-of-the-art review of integrating medical knowledge and machine learning at each stage of applying RL in the context of DTR. This section has two main objectives: first, to define how medical knowledge can be integrated into RL algorithms for developing treatment decision rules; and second, to propose adjustments to algorithms to better tailor them for DTR applications. These objectives aim to enhance the safety, interpretability, and relevance of medical decision-making tools.

4.2.1 Medical knowledge and model preparation

Like any machine learning method, the search for the optimal DTR depends on the data from which the method was trained. Data preparation is therefore an essential step. Medical knowledge is certainly involved in this process. Indeed, in this causal context, the choice of variables to collect and the selection of confounding factors are crucial. These decisions are primarily guided by medical expertise, drawn from the experience of practitioners and medical literature, as detailed in Section 3.3.3. The construction of the training dataset is thus the very first intervention of medical knowledge in RL models. It is primarily a methodological consideration that may bias the constructed optimal decision rule (Remark 3.2.9).

The second step in the preparation phase of applying RL in the context of searching for optimal DTRs involves selecting an algorithm from the various possibilities presented in Figure 3.3. This choice is primarily based on how the data were collected, the chronology of events, juxtaposed with the different characteristics of RL algorithms discussed in Section 3.3.5. The choice of method thus depends mainly on the application context and available data, and therefore, on underlying medical knowledge.

Again, this is primarily a methodological issue, where the medical specialist collaborates with the machine learning specialist to make this choice or develop a new ad hoc method. This discussion could follow the decision tree outlined in the figure titled "Overview of the guideline for the application of RL to healthcare" in [16].

4.2.2 Medical knowledge and rewards

One crucial aspect of learning optimal strategies is the formulation of rewards. This is a key component and one of the primary mechanisms for integrating medical knowledge into RL methods. In practical terms, commonly, the choice of reward is directly based on medical expertise. It is primarily a methodological issue closely linked to the definition of the study's objective. The selection of the reward is similar to choosing the primary outcome in the design of a clinical trial, with the same imperatives of precision and representativeness of the variable. Rewards mainly consist in scores or quantitative variables, such as changes in body mass index (BMI) in weight loss studies [61], or survival functions in critical care settings [101]. Additionally, more complex rewards can be found, such as compromises or combinations of variables, as seen in oncology contexts [144], where the reward is evaluated considering tumor size, treatment toxicity, patient well-being, and survival rates. In Table 3.1, an illustration of various reward functions is provided, each aiming to achieve a specific medical objective.

It is evident that selecting an ad hoc reward for the problem under study can entail choices that are either too arbitrary or too context-specific, potentially leading to overly restrictive learning objectives. An alternative approach is to replace this choice of reward with reward shaping. Several approaches have been developed in this direction.

One way to generalize and automatically construct rewards is through inverse reinforcement learning. This method uses patient trajectories generated with expert medical decision-making to extract an estimate of the underlying reward function for these choices. Thus, it also seeks to highlight the characteristics that should be considered for its formulation. The latent medical knowledge will then be encapsulated in the estimation of the reward function. This approach has been used in the context of alcohol addiction management [104] for the search for a personalized decision-making rule. The application of inverse reinforcement learning to the framework of DTRs is also explored in the article [68], where the objective of this study is to construct a reward function as a linear combination of covariates. Inverse reinforcement learning allows for the determination of rewards from data, thereby accelerating the learning of a decision rule compared to manually constructed rewards. It is important to note that these methods assume that the physicians who generated the training data made decisions aimed at maximizing the interests of each patient. Thus, the constructed rewards are sensitive not only to the quality of the data but also to medical decisions.

Another way to generalize the construction of rewards based on expert knowledge is preference learning. A subfield of research in machine learning, it relies on the idea that the expert provides preferences between two elements, which induces a ranking among these elements. Combined with reinforcement learning, preference learning uses

this induced ranking to guide the policy learning. In a model-based and online framework [27], preference learning replaces rewards to induce a preferred action based on preferences between trajectories, states, or policies. The principle is to use a simulation model to generate all possible trajectories from all possible actions, then select the preferred ones using a preference model. In an online, model-free, and off-policy framework, learning an optimal strategy is done in three steps [2]. First, an exploratory phase where trajectories are generated by a behavior policy. Second, an expert provides preferences, which induces a ranking. Third, the model learns an optimal strategy by solving a constrained optimization problem where the preferences are modeled within the constraints. In an offline framework, preference learning separately learns the rewards and the optimal strategy. The comparisons are then used in a probabilistic model, such as the Bradley-Terry algorithm, to construct rewards by maximum likelihood estimation or neural networks [105]. These rewards are then integrated into RL algorithms. Preference learning methods, described as model-based/on-policy by [27] and model-free/off-policy by [2], use preferences on trajectories on simulated data similar to the generic cancer scenario described by [144]. In [27], patient trajectories are compared using a partial order relation that considers survival, maximum toxicity over time, and final tumor size. Meanwhile, [2] formulate expert preferences by prioritizing trajectories with superior final outcomes, which include minimal tumor size and reduced toxicity levels. Preference Learning enables the construction of rewards based on expert preferences on trajectories, allowing learning to rely on explainable choices. However, the applications described in the articles [27, 2] are based on simulated cancer data and simulated preferences and have been developed in an online framework, which is not suitable for direct clinical application. An offline, off-policy solution is proposed in [105], but it has been developed in the context of robotic or video game applications.

Other methods for constructing rewards exist, such as human-centered reinforcement learning, which utilizes rewards directly provided by an expert. The agent interprets expert feedback as numerical rewards. These approaches are detailed in [56], but they are generally applied in an online and on-policy context, which involves direct interaction of the agent with patients, thus raising ethical concerns and requiring a specific application framework beyond the scope of this article.

4.2.3 Medical knowledge and value functions

The evaluation or estimation of value functions V_n^π and Q_n^π is also a key concept in RL. In the medical context, due to the complexity of environments and the volume of available data, these assessments often suffer from a lack of precision. Integrating medical expertise can be considered to improve results.

This is particularly true when medical expertise translates into knowledge of treatment response mechanisms. Indeed, these observations can then be integrated into RL methods to guide the learning of the optimal strategy. From a technical standpoint, it is conceivable to penalize the value function: decrease the value function when mechanisms identified by an expert indicate that the treatment is inappropriate and increase the value function when the treatment is considered relevant. Actions associated with

a lower value function are less likely to be selected than those associated with a higher value function. This approach thus highlights actions considered more relevant by the expert and guides learning in the right direction. This approach was implemented for patients with renal failure in [29]. Medical experts identified that patients who do not respond to standard treatment require higher doses. The authors constructed a DTR by incorporating this clinical fact into a Q-learning algorithm. When a patient does not respond to a treatment dose, the Q-values of lower doses are penalized, thus favoring higher doses. This approach offers the advantage of reducing the need for exploration and hence the learning time. However, it was developed in an online framework using simulated data, limiting its applicability to observational data.

The integration of medical expertise can also occur through relay collaboration. The principle involves considering two concurrent value functions: Q, the usual value function, and Q^{clin}, the value function under the practitioner's strategy in a given situation. The latter comes into play only when the patient is in a critical state, as evidenced by their vital signs. Subsequently, this decision and the patient's response to treatment will be used to enrich the learning model through an enhanced value function, denoted as Q^+. Thus, the strategy for updating the value functions involves recommending treatments suggested by the RL model while seeking the expertise of physicians when the patient's condition is deemed critical. Q^+ can therefore be formalized as:

$$Q^+(s_t, a_t^+) = \begin{cases} Q^{clin}(s_t, a_t^{clin}) & \text{If the patient's covariates indicate a critical state} \\ Q(s_t, a_t) & \text{Otherwise} \end{cases}$$

where a^{clin} is the treatment chosen by the clinician.

This approach has been deployed in the context of intensive care treatment in [128] when the patient exhibits severe symptoms. In such situations, RL algorithms may propose aggressive treatment strategies to maximize reward, which can entail significant risk for the patient. In this study, a model based on value functions Q incorporates human expertise on the treatment of sepsis. Applied to the MIMIC database, this model is evaluated using a score reflecting the patient's critical state. Expert intervention is triggered when the score is considered low. The application of this method demonstrates a higher survival rate compared to some similar methods without human expertise and also improves the estimation of the value function.

The principle of collaboration between the agent and the expert is also addressed in the article [109] using the MIMIC database. It still impacts the Q-functions, but now through a statistical test. The idea is to introduce exploration into an offline model by comparing risks between two strategies. One simulates standard medical decisions, while the other strategy suggests an alternative treatment. From a comparison test on state values associated with a policy, one of these strategies is adopted. The question is: when could a new treatment be better than conventional therapies? The solution seeks to balance choices of standard treatments with new options while assessing risks to discover promising alternatives that physicians have not considered.

This connection between RL and medical expertise allows for both the supervision of treatments in complex cases and the exploration of alternative treatments while assessing associated risks. Although off-policy RL can be subject to data biases, these

methods offer the potential to improve medical practice by combining data-driven insights with physicians' perspectives. Indeed, integrating medical knowledge relies primarily on observing medical mechanisms from health data or on direct input from a physician. In both cases, this integration must balance between data, expert opinion, and statistical models to determine the most suitable treatment strategies. This issue is part of a broader research question: What is the relative impact of expert knowledge, data, and machine learning agents?

4.2.4 Medical knowledge and objective function

Value-based approaches can benefit from the integration of medical expertise in determining optimal strategies. Similarly, methods for incorporating medical expertise have been proposed for policy-based approaches, which directly modify on the objective function.

Supervised reinforcement learning merges two subfields of machine learning: supervised learning and RL. The fundamental principle of this method is to maximize a long-term objective, with the supervision of an expert, in order to maintain consistency with clinical treatment standards. Its ultimate goal is to predict an optimal treatment policy, minimizing deviations from medical expert recommendations. In this framework, the expert plays a crucial role as a reference for training the RL algorithm, using a database containing all medical decisions made within a cohort. This control affects the objective function in two ways. The latter is simplified into two parts: the first, derived from an actor-critic algorithm, aims to perfectly mimic the experts through its "critic" part (Section 3.3.5.2). The second part of supervised learning minimizes the difference between predicted treatments and those traditionally administered. This method, described notably in [134], is applied in the intensive care domain using the MIMIC database and focusing on ventilation and sedation dosing. The primary objective is to provide optimal care that respects both short-term and long-term goals for patients, while adhering to best clinical practices. In this context, research shows that the supervised reinforcement learning approach outperforms the classical Actor-Critic approach in terms of convergence speed and alignment with usual medical decisions. In the study by [124], the supervised reinforcement learning approach was applied to the MIMIC dataset. The treatment recommendations obtained would lead to a decrease in patient mortality rates. Supervised reinforcement learning, in its fundamental construction, aims to perfectly mimic the usual treatment practices, making it an excellent means of emulating practitioners. However, it prevents for the proposal of alternative or less explored treatments compared to usual care methods.

4.2.5 Medical knowledge and policy

It is important to note that medical decision rules constructed within the framework of reinforcement learning recommend only a single action for a given state. The multiple policies approach involves proposing different equivalents or closely related strategies for a given patient state. Consequently, the specialist, relying on their exper-

tise and the constraints of their environment, chooses the treatment from the selection of actions offered. This approach introduces the notion of quasi-equivalent actions that may take into account considerations such as side effects, less invasive treatments, and local availability. Essentially, the general idea is to train a set of policies evaluated by value functions, which learn a correspondence between each state and a collection of closely comparable actions. Subsequently, the approach involves restricting the choice of actions by evaluating the extent to which the deviation from optimal value is acceptable. This is the concept of worst-case value, referring to the expected gain in the worst possible scenario within the set of allowed actions. The level of deviation from optimality allowed will be controlled by a hyperparameter.

The concept of multiple policies was introduced in [70] and applied in a simulated setting of sequential clinical trials for patients suffering from depression. It was developed within a model-based, on-policy, online framework with a finite horizon, not conducive to observational data or real clinical applications. In the article [114], the method evolved into a model-free and off-policy framework, still online using the temporal difference learning algorithm, and was applied in the simulated context of critical care based on MIMIC. Like the previous method, its development in an online environment does not align with our application context, but it establishes the foundation for a model-free approach, thus representing progress towards a model suitable for DTR.

In conclusion, the concept of multiple policies has also been employed in a multi-objective context, not based on expert opinion but on patient preferences, as detailed in [65]. By combining the notion of equivalent strategies with a multi-objective framework and Pareto dominance, and considering the preferences of patients, less restrictive solutions can be obtained. This approach, applied in the CATIE study specifically tailored to the DTR context, offers decision-makers increased choice by a larger class of optimal policies. These could provide the basis for an application that integrates experts' preferences and medical knowledge, thus addressing the issue outlined in this work.

4.3 Rewards construction based on preference learning

Integrating medical knowledge is crucial for developing models that are both reliable and well suited to the inputs of the medical field. This dissertation embraces this approach by exploring the construction of rewards through preference learning, specifically targeting offline and off-policy RL models, which provide an ideal context for applying DTR to observational data. It is with this objective that our method of constructing rewards through preference learning was developed. It offers a potential solution in the context of generalizing reward construction. These methods have been applied to two simulated medical applications: the treatment of patients with obesity and the treatment of patients with cancer.

In the first example, we aim to illustrate that reward models constructed through preference learning align well with traditional reward models. The goal is to demonstrate that preference learning models exhibit similar trends and provide a comparable

quantitative score for each patient as traditional methods.

The second application serves to show that even if the rewards generated through preference learning do not perfectly replicate traditional rewards, they can still facilitate the learning of a treatment strategy that is both coherent and optimal for the study's objectives.

4.3.1 Preference learning

One of the key elements in RL that guides the learning of the optimal strategy is the rewards describing the model. As mentioned in Section 4.2.2, these rewards are a crucial aspect for integrating medical knowledge. Among the emerging methods, our work has particularly focused on the in-depth study of preference learning, with the aim of adapting it to DTR applications.

Preference learning is a subfield of machine learning research. It relies on observing or collecting preferences to induce a ranking among the elements being compared. When combined with RL, it affects the reward design. The ranking information is then used to guide the policy learning. In an offline and off-policy setting, this process occurs in three steps:

— an expert expresses preferences between pairs of elements, which induces a ranking among all the instances in the previously collected dataset.
— rewards are constructed using a Bradley-Terry probabilistic model.
— these rewards are used for learning the policy in backward Q-learning models (see Section 3.2.4.2).

In the context of learning from longitudinal medical data, the data are generated from a sequence of medical decisions or a clinical trial protocol. The initial step involves collecting preferences or defining a preference rule in collaboration with an expert. Preferences can be expressed regarding patient states s_n, patient trajectories τ_n, or treatment strategies π. In this study, we focused on two specific types of preferences: first, a preference rule for comparing the conditions of different patients, and second, a rule for comparing the overall trajectories of patients. We did not consider comparisons between patient histories $h_n = (s_0, a_0, \ldots, a_{N-1}, s_n)$, as our rules do not incorporate preferences related to treatments. Concrete examples of these preference rules will be provided for two applications: one concerning the treatment of adolescents with obesity [8, 61], and the other involving a generic cancer simulation [144].

The comparisons between patient i and patient j are captured in the set $\mathcal{D}$, defined as $\mathcal{D} = \{(i, j, k_{ij}) \mid i, j \in \{1, 2, \ldots, q\}, k_{ij} \in \mathbb{N}\}$. In this set, each element (i, j, k_{ij}) represents a comparison between patient i and patient j, where i and j range from 1 to q (with q being the total number of patients). The preference score k_{ij} is a natural number that indicates the degree of preference for patient i over patient j. This set of preferences can be constructed for each individual step of treatment or for a complete trajectory, and it may also represent a global comparison of the entire patient care pathway.

From the collected preferences, the second step is to estimate the rewards using a probabilistic model. We chose to use the Bradley-Terry model, as proposed in [105], due

to its demonstrated advantages in reward construction. This model, introduced in the 1950s [10], is particularly used for ranking problems or comparisons where elements need to be ordered based on their relative performance. The Bradley-Terry model, when applied to reward construction considering preferences such as $i > j$, is presented as follows:

$$\mathbb{P}(i > j) = \frac{R_i}{R_i + R_j}$$

where $R_i = e^{\beta_i}$ is the positive real score parameterized by β associated with individual i, and will be considered as the reward associated with patient i in the RL application.

The article [105] uses a neural network model for the parametric estimation of R_i. We opted for a simpler approach based on maximum likelihood estimation. This method offers several advantages: it requires fewer data, has reduced computational complexity, and allows for a more straightforward implementation. Thus, to estimate R_i using maximum likelihood, we introduce w_{ij} as the number of times i was preferred over j. The likelihood of $R_1, \ldots, R_k$, where $k \in \mathbb{N}$ is the number of patients, is given by

$$L(R) = \prod_{(i,j)\in\mathcal{D}} (\mathbb{P}(i > j))^{w_{ij}} = \prod_{(i,j)\in\mathcal{D}} \left(\frac{R_i}{R_i + R_j}\right)^{w_{ij}}$$

The corresponding log-likelihood is then:

$$\begin{aligned} l(R) &= \ln L(R) \\ &= \sum_{(i,j)\in\mathcal{D}} \ln \left(\frac{R_i}{R_i + R_j}\right)^{w_{ij}} \\ &= \sum_{(i,j)\in\mathcal{D}} w_{ij} \ln \left(\frac{R_i}{R_i + R_j}\right) \\ &= \sum_{(i,j)\in\mathcal{D}} w_{ij} \left(\ln R_i - \ln(R_i + R_j)\right) \end{aligned}$$

The partial derivative of the log-likelihood can be decomposed into two distinct terms. First:

$$\frac{\partial}{\partial R_i} \sum_{(i,j)\in\mathcal{D}} w_{ij} \ln R_i = \sum_{(i\neq j)\in\mathcal{D}} \frac{w_{ij}}{R_i}$$

The second term is obtained by applying the chain rule:

— $\frac{\partial}{\partial R_i} \sum_{(i,j)\in\mathcal{D}} w_{ij} \ln(R_i + R_j) = \sum_{(i\neq j)\in\mathcal{D}} w_{ij} \frac{1}{R_i+R_j} \frac{\partial(R_i+R_j)}{\partial R_i} = \sum_{(i\neq j)\in\mathcal{D}} \frac{w_{ij}}{R_i+R_j}$

— $\frac{\partial}{\partial R_j} \sum_{(i,j)\in\mathcal{D}} w_{ij} \ln(R_i + R_j) = \sum_{(j\neq i)\in\mathcal{D}} \frac{w_{ji}}{R_i+R_j}$

Thus, we obtain:

$$\frac{\partial}{\partial R_i} \sum_{(i,j)\in\mathcal{D}} w_{ij} \ln(R_i + R_j) = \sum_{(i\neq j)\in\mathcal{D}} \frac{w_{ij} + w_{ji}}{R_i + R_j}$$

The final partial derivative obtained is:

$$\frac{\partial l(R)}{\partial R_i} = \sum_{(i \neq j) \in \mathcal{D}} \frac{w_{ij}}{R_i} - \frac{w_{ij} + w_{ji}}{R_i + R_j}$$

The preference rules on which we base the learning of our rewards compare each pair of individuals only once. Consequently, we have $w_{ij} = 1$ for each pair (i, j), which can lead to a database considered weak in terms of information. To address this limitation, it is common in the literature on comparison models to use variants of the Bradley-Terry model with Lasso or Ridge penalization [122, 120, 25].

The penalized log-likelihood for the Bradley-Terry model with L2 regularization (Ridge), which we chose to apply, is given by:

$$\begin{aligned} l_{\text{pen}}(R) &= \sum_{(i,j) \in \mathcal{D}} w_{ij} \ln \left(\frac{R_i}{R_i + R_j} \right) - \lambda \sum_i R_i^2 \\ &= \sum_{(i,j) \in \mathcal{D}} w_{ij} \left(\ln R_i - \ln(R_i + R_j) \right) - \lambda \sum_i R_i^2 \end{aligned}$$

where λ is the regularization parameter that controls the strength of the penalty applied to the rewards.

The estimation of R_i is performed using the Newton-Raphson optimization algorithm. This algorithm adjusts the parameters at each iteration based on a gradient descent approach.

When the model is based on preferences between states, we construct a reward $r(S_t, S_{t+1})$ for each state transition of a patient. In this case, the classic form of backward Q-learning can be applied using an induction regression model. The Q-function, $\hat{Q}_N$, at the final step N is obtained by regressing the values of h_N on r_N. The Q-functions, $\hat{Q}_n$, for the previous steps are obtained by regressing the values of h_n on $r_n + \max_{a_n \in \mathbb{A}(s_n)} \hat{Q}_{n+1}(s_n, a_n)$.

When the model is based on preferences between trajectories, we construct a single final reward for each patient. In this case, the classic backward Q-learning algorithm, which relies on intermediate rewards, can no longer be used. Instead, we opt for a regression method that does not depend on intermediate evaluations or rewards [79, 53, 61]. The estimation of the final Q-function, $\hat{Q}_N$, remains unchanged. However, the Q-functions, $\hat{Q}_n$, are obtained by regressing the values of h_n on $\max_{a_n \in \mathbb{A}(s_n)} \hat{Q}_{n+1}(s_n, a_n)$.

4.3.2 BMI data application

4.3.2.1 Data

The `bmiData` dataset contains simulation data that mimics a two-stage clinical trial, similar to the one described in [8]. The goal of this study was to reduce the Body Mass Index (BMI) in adolescents with obesity by personalizing the treatment at each stage. The treatments considered are meal replacement and conventional diet, coded as $\{-1, 1\}$ respectively. It includes data from 210 patients.

This dataset contains the following information:
— The adolescent's gender: **gender**
— The adolescent's race: **race**
— Parents' BMI: **parent BMI**
— BMI at the start of the study: **baseline BMI**
— The first treatment chosen: **A1**
— BMI at the fourth month: **month 4 BMI**
— The second treatment administered: **A2**
— BMI at the twelfth month of the study: **month 12 BMI**

The `bmiData` dataset was used in [61] to illustrate the functioning of backward Q-learning based on a linear regression model. It is available in two R packages: **iqlearning** [61] and **DynTxRegime** [1].

In the context of SMART studies, it is common to evaluate the progression of a patient's treatment based on a final outcome variable. In this study, we consider the change in BMI as the final response, defined by:

$$R_2 = -100 \times \frac{\text{bmiData['month 12 BMI']} - \text{bmiData['baseline BMI']}}{\text{bmiData['baseline BMI']}}$$

When the RL model requires an intermediate response, we will use the change in BMI after 4 months, calculated as follows:

$$R_1 = -100 \times \frac{\text{bmiData['month 4 BMI']} - \text{bmiData['baseline BMI']}}{\text{bmiData['baseline BMI']}}$$

4.3.2.2 Preference rules

We generated two preference rules: one based on each state s_n and the other on trajectories τ. Patients are compared pairwise according to these rules before applying a reward estimation model based on the Bradley-Terry model.

Preference rule based on states At step $t = 1$, the preference between patients is determined by comparing the reduction in BMI after 4 months. Patient i is preferred over patient j if the weight loss observed for i after 4 months is greater than that for j. In other words, if the reduction in BMI for i compared to its initial value is greater than that of j, then i is considered to have a better performance at this stage. In all other cases, the patients are deemed incomparable, and the score assigned is 0.

At step $t = 2$, preferences are evaluated based on two distinct levels of criteria. First, the variation in BMI between the fourth and twelfth months is compared. A patient is preferred if their weight loss during this period is more significant. Thus, the patient with the greatest reduction in BMI between these two dates receives a score of 2. Second, if the weight changes between the fourth and twelfth months are equivalent, the total BMI loss from the start to the end of the study is examined. The patient with the greatest total BMI loss is preferred, and a score of 1 is assigned to

1. Available on CRAN: https://cran.r-project.org/web/packages/DynTxRegime/index.html.

the preferred patient. In summary, this method prioritizes the comparison criteria: the first level focuses on recent weight loss (between the 4th month and the 12th month of the study), while the second level evaluates the total weight loss since the beginning of the study. This approach provides a clear preference between patients based on their improvement over time. In all other cases, the patients are deemed incomparable, and the score assigned is 0.

With this comparison rule, we will obtain, for each step, an estimated reward vector, denoted $\mathbf{R}_1^{BT}$ and $\mathbf{R}_2^{BT}$. These vectors will be compared to the manual reward vectors $\mathbf{R}_1$ and $\mathbf{R}_2$.

Remark 4.3.1. In our preference models, we introduce preference levels. Assigning a score of 2 indicates that this situation is preferred over one assigning a score of 1. This does not imply that it is twice as good, but it allows us to establish ordered relationships between patients.

Preference rules based on trajectories To evaluate preferences between patients based on their trajectories, we use a hierarchical method based on several levels of criteria. Firstly, we compare weight loss continuously throughout the study. A patient is considered to have a superior overall performance if their weight loss is more significant at each stage compared to another patient. This approach, reflecting consistent improvement over time, is assigned a score of 2. Secondly, if these comparisons do not provide a clear distinction, we analyze the overall change in BMI from the beginning to the end of the study. A patient is preferred if they have lost more weight over the entire period, even if this loss was not progressive. This criterion is associated with a score of 1. Thus, this hierarchical comparison method prioritizes criteria by first giving preference based on progressive weight loss throughout the study, followed by overall weight loss. This allows for establishing preferences between patients based on their weight loss trajectories.

In all other cases, patients are not comparable, and the score assigned is 0.

With this comparison rule, we will obtain an estimated reward vector, denoted $\mathbf{R}_T^{BT}$. This vector will be compared to the manual response studied when there is no intermediate response, $\mathbf{R}_2$.

4.3.2.3 Results

The first objective is to compare the rewards constructed traditionally with those generated through the preference model. To facilitate this comparison and place the rewards on a common scale, they have been standardized. We will start with a descriptive study of the two types of rewards. Next, we will perform a correlation analysis to assess the relationship between these rewards. Lastly, we will examine the final results in terms of variable importance after applying backward Q-learning based on linear regression.

Descriptive statistics The histograms presented in Figures 4.1 and 4.2 reveal a similar distribution of the data for the classical rewards in blue and those generated by

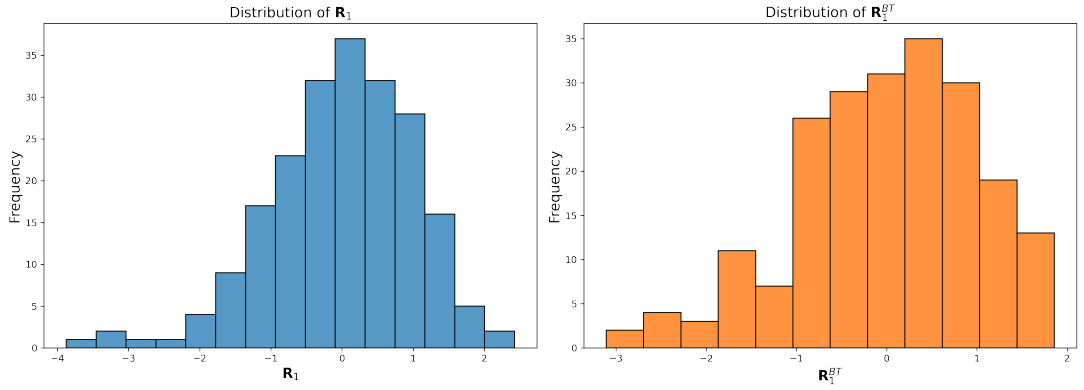

Figure 4.1 – Comparative histograms of reward models in the first stage

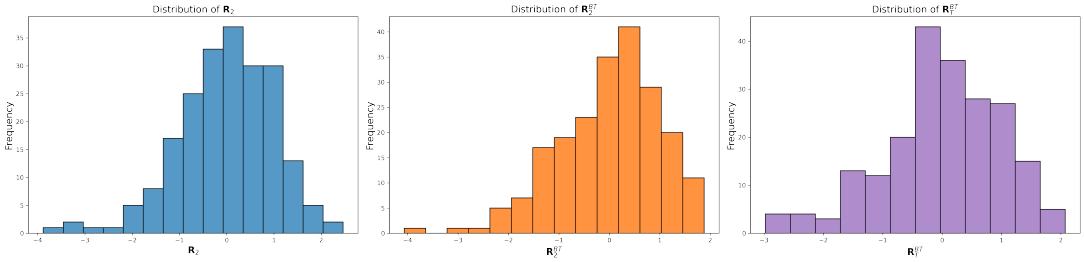

Figure 4.2 – Comparative histograms of reward models in the second stage

the preference model, with rewards by stage shown in orange and rewards by trajectory shown in purple. This comparison suggests that the different types of rewards share comparable distribution characteristics. The analysis of the boxplots shown in Figures 4.3 and 4.4 further confirms this similarity. Indeed, the boxplots display reward distributions that, while potentially exhibiting specific variations, show comparable overall structures. These observations reinforce the idea that both the classical rewards and those derived from the preference model share similar dispersion characteristics.

The analysis of Figure 4.5 shows a positive, linear trend between the classical rewards and those generated by the preference model. This linear relationship suggests that, despite the different calculation methods, the two types of rewards are closely related and follow a similar trend.

Correlation analysis To further explore the correlation between the reward models, we calculated several metrics. We first computed the Pearson correlation coefficient to measure the strength and direction of the linear relationship between the rewards from different models. The results indicate a moderately strong positive correlation. Next, we used the Spearman rank correlation coefficient to check if the rewards from the models change together in a monotonous way, meaning they follow a consistent trend, even if it is not linear. This also showed a moderately strong correlation. Finally, we used Kendall's coefficient to assess the level of correlation between the reward models.

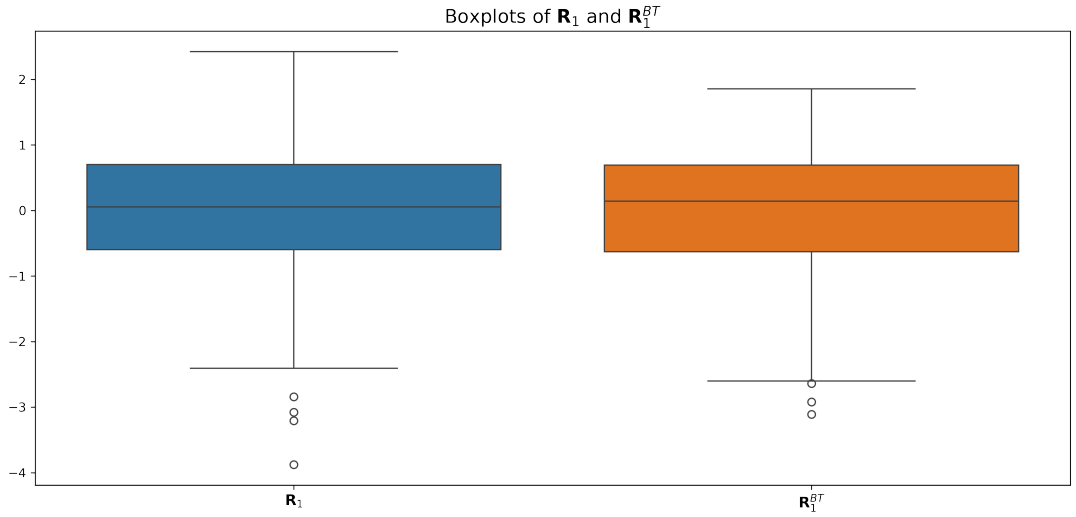

Figure 4.3 – Comparative boxplots of reward models in the first stage

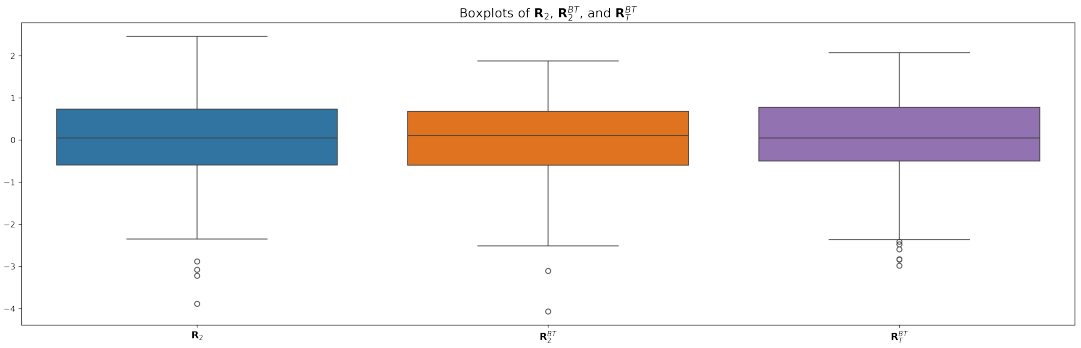

Figure 4.4 – Comparative boxplots of reward models in the second stage

This coefficient compares how often the order of rewards is the same or different between models. A high Kendall's coefficient indicates a strong correlation, while a low coefficient suggests less correlation. The results show a moderate correlation between the classical rewards and those generated by the preference model. The associated p-values are very low, indicating that these correlations are statistically significant and unlikely to be due to chance.

Importance features For the stage-based preference models, as discussed at the end of Section 4.3.1, we used the Q-learning method described in Chapter 2, Section 3.2.4.2. We chose support vector regression with a linear kernel as our regression model. This approach allows us to determine the importance of features for each model trained with different rewards. As shown in Table 4.2, the signs of the coefficients are the same for all features, except for the gender and race of patients at the first stage. The magnitudes are similar, and the ranking of feature importance is almost entirely preserved.

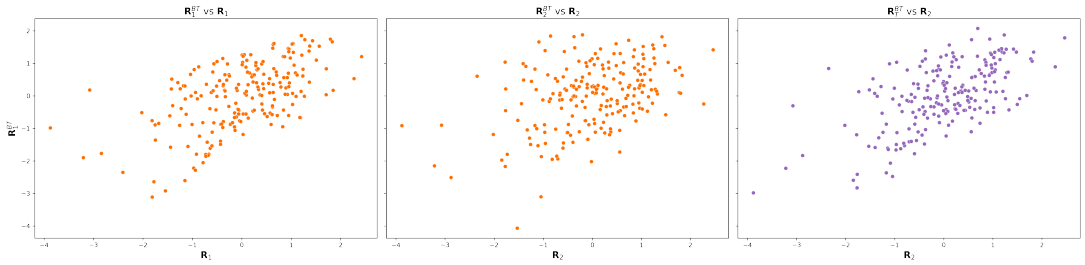

Figure 4.5 – Scatter-plots between traditional rewards and preference-based rewards

	Correlation coefficients study		
	$\mathbf{R}_1/\mathbf{R}_1^{BT}$	$\mathbf{R}_2/\mathbf{R}_2^{BT}$	$\mathbf{R}_2/\mathbf{R}_T^{BT}$
Pearson	0.57 [0.47, 0.65]	0.41 [0.29, 0.51]	0.62 [0.52, 0.70]
Spearman	0.55 [0.45, 0.64]	0.35 [0.23, 0.46]	0.56 [0.45, 0.66]
Kendall	0.39 [0.31, 0.46] 7×10^{-17}	0.24 [0.15, 0.33] 2×10^{-7}	0.40 [0.32, 0.48] 3×10^{-18}

Table 4.1 – Correlation coefficients between $\mathbf{R}_1/\mathbf{R}_1^{BT}$, $\mathbf{R}_2/\mathbf{R}_2^{BT}$, and $\mathbf{R}_2/\mathbf{R}_T^{BT}$.The 95% confidence intervals are shown in brackets. For Kendall, the first line represents the correlation coefficient, and the second line gives the p-value from the Kendall rank correlation test. This test checks if the observed association between two variables is significantly different from what would be expected under the null hypothesis of no association.

For the trajectory-based preference models, as mentioned at the end of Section 4.3.1, we used a Q-learning model without intermediate rewards. We replicated the Q-learning model from `bmiData` in [61], which is based on linear regressions with interactions between treatments and state variables. The results, shown in Table 4.3, lead to the same conclusions as before.

Remark 4.3.2. Constructing confidence intervals for the results in Tables 4.2 and 4.3 is challenging due to the non-differentiable nature of the regression models used in Q-learning (i.e., maximization). In [54], the authors discuss interval estimation and propose a locally consistent confidence interval for parameters indexing the optimal DTR.

	First Stage		Second Stage	
	$\mathbf{R}_1$	$\mathbf{R}_1^{BT}$	$\mathbf{R}_2$	$\mathbf{R}_2^{BT}$
Gender	0.03	-0.20	-0.01	-0.12
Race	-0.12	0.22	0.02	0.10
Parent_BMI	-1.28	-0.69	-0.03	-0.07
Baseline_BMI	0.83	0.11	1.19	0.48
A1	-	-	0.00	0.03
Month4_BMI	-	-	-1.37	-0.76

Table 4.2 – Q-learning coefficients, indicating the influence of each variable in the regression model, for the first and second stages, with manual rewards and stage preference learning rewards.

	First Stage		Second Stage	
	$\mathbf{R}_2$	$\mathbf{R}_T^{BT}$	$\mathbf{R}_2$	$\mathbf{R}_T^{BT}$
Gender	-0.05	0.00	-0.04	0.00
Race	0.00	0.01	-0.12	-0.11
Parent_BMI	-0.19	-0.08	-0.12	-0.15
Baseline_BMI	-0.04	0.01	-0.44	-0.15
A1	-0.04	0.01	0.05	0.10
A1:Gender	0.02	0.01	0.02	0.01
A1:Parent_BMI	-0.11	-0.01	0.02	0.13
Month4_BMI	-	-	-0.44	-0.15
A2	-	-	0.05	0.10
A2:Parent_BMI	-	-	0.15	0.13
A2:Month4_BMI	-	-	0.02	0.16

Table 4.3 – Q-learning coefficients, indicating the influence of each variable in the regression model, for the first and second stages with manual rewards and trajectory preference learning rewards.

4.3.3 Cancer application

4.3.3.1 Data

In this second application case, we consider a simulation of non-specific cancer treated with chemotherapy, based on a model proposed by [144]. This model is frequently used as a case study in RL [27, 2, 41, 32]. It relies on four principles:

1. **Tumor growth without chemotherapy** : the model simulates the natural progression of the tumor if no treatment is administered.
2. **Negative effects of chemotherapy on patient well-being** : the side effects of chemotherapy are modeled, reflecting their impact on the patient's quality of life.
3. **Drug efficacy against tumor cells and increased toxicity** : The model

takes into account the drug's ability to eliminate tumor cells while increasing toxicity for the patient.

4. **Interaction between tumor cells and patient well-being** : a dynamic interaction between tumor progression and the impact on patient well-being is integrated into the model.

For each patient, there are two state variables $S_t = \{Y_t, X_t\}$, where Y represents the tumor size and X represents the toxicity of the treatment at each month t such that $t = 0, \ldots, 6$. The treatment A_t administered at month t is a dosage between 0 and 1 with a step of 0.1. This model is based on the following system of differential equations:

$$\begin{aligned}\Delta Y_t &= [0,15 \times \max(X_t, X_0) - 1,2 \times (A_t - 0,5)] \times \mathbb{1}(Y_t > 0)\\ \Delta X_t &= 0,1 \times \max(Y_t, Y_0) + 1,2 \times (A_t - 0,5)\end{aligned}$$

By using the indicator function $\mathbb{1}(Y_t > 0)$, the model assigns the status of complete remission to a patient when the size of their tumor is reduced to zero, indicating no recurrence.

The possibility of a patient's death during a treatment is represented by a survival model. For each time interval $(t-1, t]$, the survival rate is defined as a function of the tumor size and toxicity: $\lambda(t) = \exp(-4 + Y_t + X_t)$. In this model, both tumor size and toxicity have an equally important influence on the patient's survival. The probability of the patient dying during the time interval $(t-1, t]$ is given by:

$$\mathbb{P}_{\text{décès}} = 1 - \exp\left(-\int_{t-1}^{t} \lambda(x)dx\right)$$

Remark 4.3.3. In a non-Markovian framework, observational applications of DTRs typically involve a limited number of steps, rarely exceeding four. However, to enable a direct comparison with the results obtained in [144], we extended the model to six stages.

Remark 4.3.4. Backward Q-learning is specifically designed to be applied to data with complete trajectories, to maintain consistency in dimensions throughout the induction process. However, [144] does not specify how to handle incomplete trajectories. To address this gap, we adopted an approach for patients who were either in remission or had died before the end of the trajectory. In these cases, the last available values for tumor size Y and toxicity X are extended to the end of the trajectory, with a chemotherapy dose fixed at zero. The survival status and remission, as well as the associated stage, are preserved as indicators for subsequent analysis.

The reward model, which we will compare to and which is proposed by [144], is defined as follows:

$$R_{t,1} = \begin{cases} -60 & \text{if the patient dies} \end{cases}$$

$$R_{t,2} = \begin{cases} 5 & \text{if } X_{t+1} - X_t \leq -0.5 \\ -5 & \text{if } X_{t+1} - X_t > -0.5 \end{cases}$$

$$R_{t,3} = \begin{cases} 15 & \text{if } Y_{t+1} = 0 \\ 5 & \text{if } Y_{t+1} - Y_t \leq -0.5 \text{ and } Y_{t+1} \neq 0 \\ -5 & \text{if } Y_{t+1} - Y_t > -0.5 \end{cases}$$

$$R_t = R_{t,1} + R_{t,2} + R_{t,3}$$

4.3.3.2 Preference rules

Preference rule based on states The preference rule between two patients at stage n is defined according to several criteria that allow comparison based on their health status. The criteria are as follows:

1. **Remission**: if patient i is in remission at stage n and their remission stage is lower than that of patient j, then patient i is preferred over patient j. In this case, patient i receives a score of 3.
2. **Death**: if patient i is still alive at stage n while patient j is deceased at this stage, patient i is preferred. Here, patient i receives a score of 2.
3. **Tumor size and toxicity**: if the tumor measurement Y_n and toxicity X_n for patient i are both lower than those for patient j at stage n, then patient i is preferred. In this case, patient i receives a score of 1.

In all other cases, patients are not comparable, and the score assigned is 0.

In summary, this model assigns preference scores first based on remission, then on death status, and finally on tumor and toxicity measurements.

Using this comparison rule, we obtain preference score vectors $\mathbf{R}_t^{\mathrm{BT}}$ for each stage transition t. These reward vectors can then be compared to the reward vectors $\mathbf{R}_t$ provided by the classical model.

Preference rule based on trajectories This preference rule, partly inspired by [27], compares two patients based on several criteria related to their health status. The criteria are defined as follows:

1. **Remission**: if one patient is in remission while the other is not, the patient in remission is preferred. This situation is associated with a score of 2.
2. **Death**: if one patient has died at a given stage while the other is still alive at that stage, the patient who is still alive is preferred. This case is also associated with a score of 2.
3. **Tumor size and toxicity**: when the previous conditions do not determine a preference, we compare patients based on the maximum toxicity and the final tumor size over their entire trajectory. Specifically:
   — the maximum observed toxicity over the trajectory must be less than or equal for the preferred patient.

— the final tumor size over the trajectory must also be less than or equal for the preferred patient.

If these conditions are met, the patient with the most favorable maximum toxicity and tumor size receives a score of 1.

In all other cases, patients are not comparable, and the score assigned is 0.

In summary, this preference model assigns scores based on remission, death, and overall toxicity and tumor size on the trajectory.

Using this comparison rule, we will obtain a reward vector $\mathbf{R}_T^{\mathrm{BT}}$ for each patient. Learning in this specific context will be performed using backward Q-learning without intermediate rewards, as specified when introducing the model.

4.3.3.3 Results

In a similar manner as previously, we will compare the traditionally constructed rewards with those generated through preference models. To facilitate this comparison and provide a comprehensive overview, we will compare the cumulative rewards for [144] model and the state-based preference model with those obtained from the trajectory-based preference model, using histograms and boxplots. The rewards have been standardized. Correlation analysis will be conducted both on cumulative rewards and on a step-by-step comparison.

The primary goal of this application is to study and compare the learned strategies based on different rewards. To evaluate them, we will observe their application on 10,000 new patients and study the average change in their tumor size, treatment toxicity, and the combination of both.

Descriptive statistics The blue histogram in Figure 4.6, corresponding to the cumulative reward from the initial model [144], clearly reflects the discretization with the possible value combinations. The orange histogram, related to the state-based preference model, and the purple histogram, related to the trajectory-based preference model, show that these methods yield continuous rewards. The distribution for each is different. This difference is also highlighted by the boxplots, with the same conclusion in terms of dispersion, in Figure 4.7. The correlation graphs in Figure 4.8 do not show linear correlation.

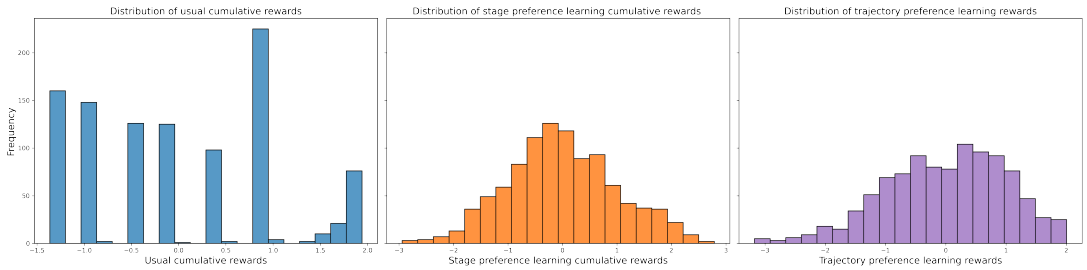

Figure 4.6 – Comparative histograms of reward models for generic cancer application

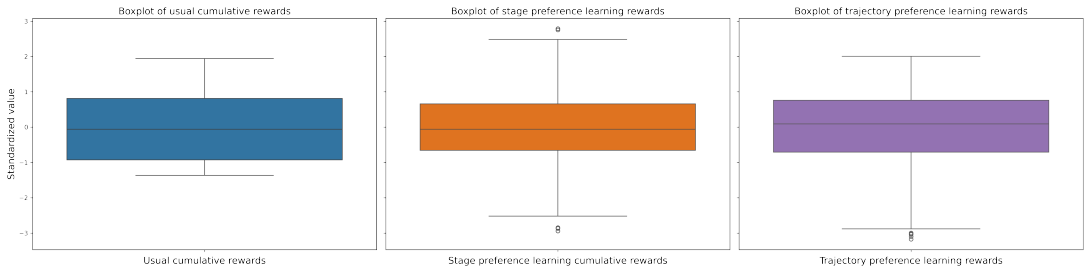

Figure 4.7 – Comparative boxplots of reward models for generic cancer application

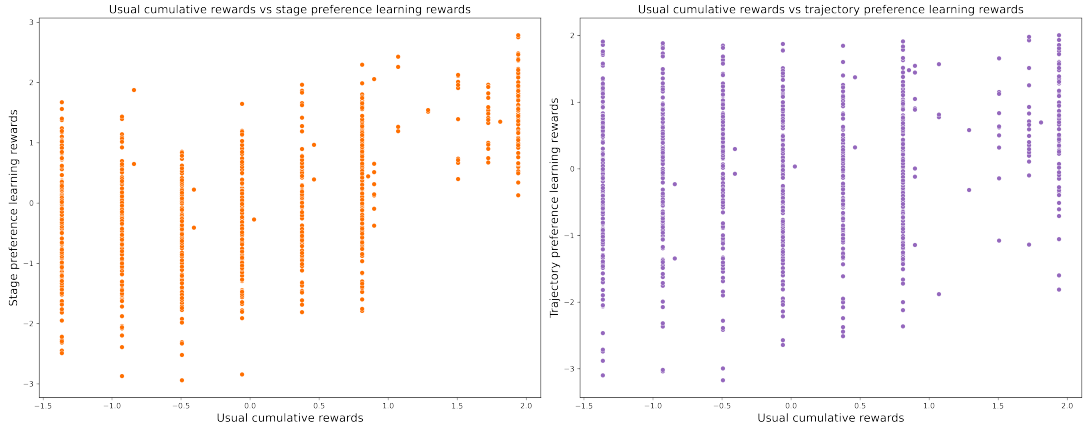

Figure 4.8 – Scatter-plots between traditional rewards and preference-based rewards for generic cancer application

Correlation analysis The Pearson and Spearman correlation coefficients presented in Table 4.4 between the traditional cumulative rewards and those obtained from the stage-based preference model are moderately strong. However, when comparing them stage by stage, as shown in Table 4.5, the correlations are weak, and very weak for the last stage. The Pearson and Spearman correlation coefficients between the cumulative rewards from the classical model and those obtained from the trajectory-based preference models are weak. The reward models obtained through preference learning are not correlated with those described in [144].

	$\sum \mathbf{R}^{Zhao} / \sum \mathbf{R}^{SPL}$	$\sum \mathbf{R}^{Zhao} / \mathbf{R}_T$
Pearson	0.55 [0.50, 0.59]	0.14 [0.08, 0.20]
Spearman	0.50 [0.44, 0.55]	0.11 [0.05, 0.17]

Table 4.4 – Correlation coefficients between the cumulative reward $\sum \mathbf{R}^{Zhao}$ from [144] and the cumulative reward of stage preference learning $\sum \mathbf{R}^{SPL}$ or the trajectory preference learning model $\mathbf{R}_T$. The 95% confidence intervals are indicated in brackets.

	$\mathbf{R}_0/\mathbf{R}_0^{BT}$	$\mathbf{R}_1/\mathbf{R}_1^{BT}$	$\mathbf{R}_2/\mathbf{R}_2^{BT}$
Pearson	0.38 [0.31, 0.43]	0.20 [0.14, 0.26]	0.24 [0.18, 0.30]
Spearman	0.39 [0.33, 0.43]	0.25 [0.19, 0.31]	0.26 [0.20, 0.32]
	$\mathbf{R}_3/\mathbf{R}_3^{BT}$	$\mathbf{R}_4/\mathbf{R}_4^{BT}$	$\mathbf{R}_5/\mathbf{R}_5^{BT}$
Pearson	0.26 [0.20, 0.32]	0.23 [0.17, 0.29]	0.02 [-0.04, 0.08]
Spearman	0.26 [0.21, 0.32]	0.23 [0.17, 0.29]	0.01 [-0.05, 0.08]

Table 4.5 – Correlation coefficients between manual rewards [144] and stage preference learning rewards at each stage. The 95% confidence intervals are indicated in brackets.

Optimal policies The objective of this study is to compare the optimal strategies generated by the different reward models. To achieve this, we examined the evolution of tumor size and treatment toxicity. All these quantities were averaged over 10,000 simulated patients based on the model presented in Section 4.3.3.1, following the different learned treatment strategies. Each model was trained on 1,000 patients. For learning strategies from rewards generated by the state-based preference model and the traditional rewards presented in the reference paper, we used classical backward Q-learning. For learning strategies from rewards constructed by the trajectory-based preference model and for the manual cumulative rewards, we used backward Q-learning without intermediate rewards. In both cases, the regression model chosen is based on Support Vector Regression. These results were also compared to the administration of constant dosages among the possible dosages $0.1, 0.2, \ldots, 1.0$.

In Figures 4.9 and 4.10, we observe the average evolution of tumor size and treatment toxicity, either with a constant treatment or by following the four different strategies. None of the proposed strategies show better results than those achieved with a constant dosage. This result is expected, as explained in [144]: "because when a higher dose level decreases tumor size, it can yield a higher toxicity simultaneously, and vice versa. However, due to our reward functions structure, the estimated optimal policies have an appealing feature that seeks a good balance between toxicity and efficacy."

Thus, Figure 4.11 illustrates the combined average results of tumor size and treatment toxicity. In the long term, the four models outperform the constant dosage treatment plans. We observe that the model based on rewards from a trajectory-based preference model achieves the lowest performance. Strategies based on the stage-based preference model or the classical model from [144] yield nearly identical results. The model with the best performance in terms of balancing toxicity and tumor size is the one derived from the cumulative rewards of the [144] model.

4.3.4 Conclusion

The objective of our first case study on the `bmiData` dataset was to demonstrate that rewards constructed using a preference model, whether based on stages or trajectories, capture the same variations and dispersion as the observed rewards in the classical approach. This result was confirmed by a descriptive study as well as an analysis of correlation coefficients. The differences observed between the two types of rewards

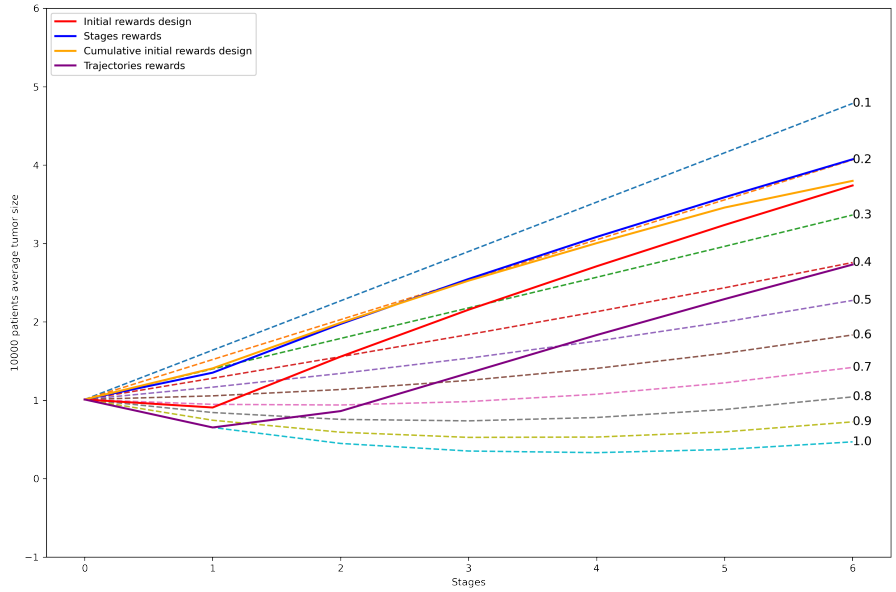

Figure 4.9 – Average tumor size across stages for different treatment policies, calculated over 10,000 patients

primarily appear in the study of feature importance. Although the results are very close, they are not perfectly identical. It is important to note that the baseline rewards perfectly fulfill their medical objective and are initially well-calibrated to address the clinical trial problem. Our method has shown its consistency in a classical application case but could be particularly advantageous in contexts where it is more challenging to manually define a reward function, especially when arbitrary variable weightings must be chosen, as in our second case study.

The objective of the second case study was to examine the strategies learned from the different models. The initial reward model, constructed manually, uses values that assess the patient's condition; these choices can be considered subjective or determined by trial and error. Notably, the decision to assign a score of -60 can be considered particularly arbitrary. The advantage of the method presented here is that it constructs rewards from a preference model, making it a more generalized approach. The reward construction is data-driven, based on rankings among all the information gathered in the database rather than on individual data. The descriptive statistical study and correlation analyses show that the rewards generated by our models and those of the initial model do not have the same distributions and are very weakly correlated. However, as shown in Figure 4.11, the performance of the strategies learned from these models produces the expected results: a medical decision rule capable of balancing the treatment's toxicity and tumor size. Moreover, the performance of our stage-based

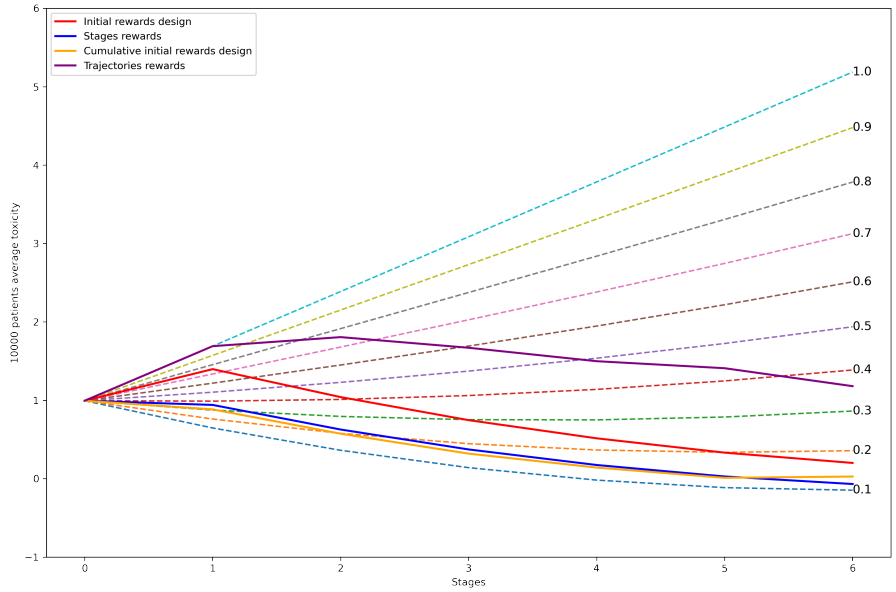

Figure 4.10 – Average toxicity across stages for different treatment policies, calculated over 10,000 patients

preference reward construction method is very similar to that obtained with the model of [144].

4.4 Perspectives

Applying RL in precision medicine requires specific adaptations. Indeed, the data generally comes from observational studies, limiting the methods to offline applications. RL algorithms for DTR must be model-free, value-based, and off-policy, as argued in the previous chapter. We evaluated all RL methods applied to DTR against these characteristics. However, when aiming for real clinical use of these methods in hospitals, a major issue emerges in the search for an optimal treatment strategy: the acceptability of the optimal DTR to both patients and practitioners. This raises concerns about how understandable the decision rules are for both patients and physicians, which is crucial for their clinical use. Integrating medical expertise into machine learning methods for personalized treatments is essential to improve safety, interpretability, and effectiveness in observational scenarios.

One way to overcome this issue is to consider algorithms involving, one way or another, medical expertise or knowledge. The integration of expert knowledge can occur at various levels in the RL application process or in its key components, such as rewards, value functions, the objective function, or the policy.

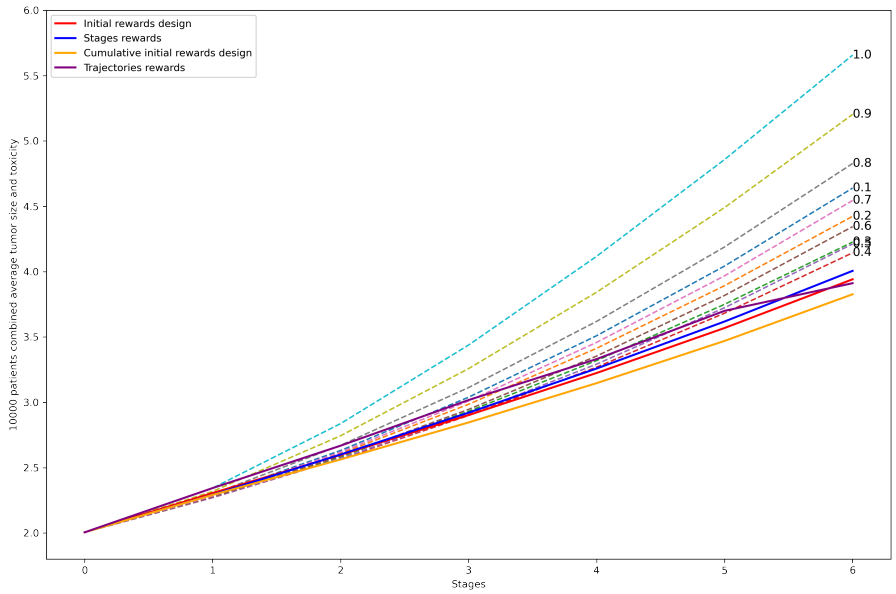

Figure 4.11 – Average toxicity combined to tumor size across stages for different treatment policies, calculated over 10,000 patients

First, the medical knowledge is often integrated before the study, at the design of the experiment. Indeed, physicians contribute to selecting the variables used for learning the decision rule. Similarly, algorithm selection involves collaboration between medical and machine learning expert, based on the application framework and available data.

Second, the medical knowledge can be integrated by acting on the rewards. Rewards is one of the main elements of a RL algorithm. Since they influence and guide the determination of the decision rule. Their design is thus crucial. Traditionally, a variable representative of the study's objective is chosen. Methods such as inverse reinforcement learning and preference learning attempt to generalize their construction through expert input. Preference learning [27, 2] and human-centered RL [56] directly incorporate expert knowledge into reward construction. However, this method suffers from being developed only in an online setup, which is not applicable to DTRs and observational clinic application. Nonetheless, early research in this area can serve as a foundation for further exploration. On the other hand, inverse reinforcement learning is promising since it is developed within the offline context and it is well suited for real clinical application [104, 68].

Thirdly, the learning of decision rules can be achieved through value functions, allowing for the integration of medical expertise at this level. One approach is to incorporate observed medical mechanisms; specifically, the idea is to penalize the Q-values associated with non-decisive treatments [29]. However, this method was initially devel-

oped in an online context and requires reassessment for offline settings. A second idea is to establish a relay between human decisions and decisions proposed by the algorithm. In one scenario, the physician would take over when the patient is in critical conditions [128]. In another scenario, the algorithm would suggest alternative treatments to those traditionally proposed, along with associated risks [109]. These hybrid methods seem promising for real clinical applications, but concrete evidence of their implementation is currently lacking. In the policy-based methodological framework, the integration of expertise can occur through a method called supervised RL [134, 124]. Its aim is to faithfully replicate common medical practices, offering precise emulation of physicians' decisions. However, it does not allow for the discovery of alternative or underexplored treatments compared to conventional care methods.

Lastly, the learning of decision rules can be approached methodologically through policy and it is worth noting that classical RL methods typically recommend only one policy, typically one treatment and one dose for each decision time. To enrich the context, multiple policies methods have been developed with the aim of offering an expert multiple equivalent treatment to choose from. The work of [65] is particularly suitable for application to observational data-based DTRs, but it was developed within a framework of patient preferences and could be reassessed within an expert preference framework.

In this manuscript, we decided to provide an initial response to the issues raised by this state of the art. We developed a method that generalizes the construction of rewards based on preference learning and can be applied to DTRs. The method we implemented consists of three steps. First, an expert expresses preferences between pairs of elements, which induces a ranking among all instances in the previously collected dataset. Second, rewards are constructed using a Bradley-Terry probabilistic model. Third, these rewards are used to learn the policy in backward Q-learning models.

In our initial case study using the `bmiData` dataset, we showed that rewards derived from a preference model, whether based on stages or trajectories, effectively capture the same variations and dispersion with traditional reward methods. This was validated by descriptive analyses and correlation studies, though some differences appeared in feature importance assessments. Although our method demonstrated reliability in a standard application, it offers significant benefits in situations where manually defining a reward function is difficult, particularly when arbitrary variable weightings are involved, as illustrated in our second case study.

The second case study, involving simulated generic cancer, aimed to evaluate the strategies derived from different models. Descriptive statistics and correlation analyses indicate that the reward distributions from our models differ significantly from those of the initial model, exhibiting very weak correlations. This expected result is due to our method's data-driven construction of rewards, which utilizes the entire dataset rather than individual data points. Consequently, as illustrated in Figure 4.11, the strategies learned from our models achieve the desired outcomes: a medical decision rule that effectively balances treatment toxicity and tumor size. Furthermore, the performance of our stage-based preference reward method closely aligns with that of the model proposed by [144].

Our work has some limitations and could potentially see performance improvements. The estimation of rewards from the Bradley-Terry pairwise comparison model relies on the Newton-Raphson algorithm. However, literature on comparison models suggests that the minorization-maximization algorithm [42] is a more powerful estimation technique. Other comparison models, from research in social choice or sports statistics, could also be considered. For instance, the Thurstone-Mosteller model [34], the Elo model [14], and the Plackett-Luce model [87] are worth mentioning. One limitation of this method is that it directly provides reward values from comparisons. It would be interesting to reformulate the models into a parametric reward function based on state variables.

The integration of medical knowledge is a promising research field, exploring various innovative perspectives and methods. However, further research is needed to adapt them to the specific constraints and realities of precision medicine. These advancements have the potential to lead to practical clinical applications and significantly enhance daily hospital operations. This aligns with the broader challenge of applying mathematical solutions effectively in clinical practice. Particularly, the development of health system science enables the use of interdisciplinary skills to study the complexity of healthcare systems [3, 46]. Practically speaking, the aim is to ease the transition of laboratory discoveries into clinical practices [31], achieved by forming interdisciplinary teams within healthcare systems. Combining progress in both research areas could establish a tangible framework for applying RL alongside medical expertise, simplifying the treatment decision process for the benefit of all involved parties. We hope this study will encourage collaboration between machine learning researchers and healthcare professionals, by showing a framework that helps practically applying RL for DTR context.

Chapter 5

Bayesian Outcome-Weighted Learning

One of the primary goals of statistical precision medicine is to learn optimal individualized treatment rules (ITRs). The classification-based, or machine learning-based, approach to estimating optimal ITRs was first introduced in outcome-weighted learning (OWL). OWL recasts the optimal ITR learning problem into a weighted classification problem, which can be solved using machine learning methods, e.g., support vector machines. In this paper, we introduce a Bayesian formulation of OWL. Starting from the OWL objective function, we generate a pseudo-likelihood which can be expressed as a scale mixture of normal distributions. A Gibbs sampling algorithm is developed to sample the posterior distribution of the parameters. In addition to providing a strategy for learning an optimal ITR, Bayesian OWL provides a natural, probabilistic approach to estimate uncertainty in ITR treatment recommendations themselves. We demonstrate our method through several simulation studies.

This research, conducted in collaboration with Nikki L. B. Freeman, is accessible on arXiv:2406.11573. At the time of writing this manuscript, it has not been submitted for publication, as we intend to augment it with an application using real clinical data.

Contents

5.1 Introduction **74**
5.2 Background **76**
5.2.1 Setting 76
5.2.2 Outcome-weighted learning 76
5.2.3 Bayesian support vector machines 77
5.3 Our approach **78**
5.3.1 Prior specification for the ITR parameters 79
5.3.2 Exponential power prior distribution for $\boldsymbol{\beta}$ 80
5.3.3 Spike-and-slab prior distribution for $\boldsymbol{\beta}$ 81
5.3.4 Estimation 81
5.3.5 Prediction and uncertainty quantification 86
5.4 Simulation studies **86**
5.4.1 Classification performance 86
5.4.2 Treatment recommendation uncertainty quantification 88
5.5 Discussion **89**
5.6 Appendix : derivation of the Gibbs sampling algorithms **92**
5.6.1 Conditional distribution of $\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, \mathbf{a}_i, \mathbf{r}_i$ 92
5.6.2 Conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\mu}_0, \sigma_0^2, \mathbf{r}, \mathbf{a}, \mathbf{x}$ (Normal prior) . . 94
5.6.3 Conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ (Exponential power prior) 99

5.1 Introduction

The task of statistical precision medicine is to learn from data how to match patients to treatments with the aim of improving health outcomes [50]. One way to operationalize this goal is through individualized treatment regimes (ITRs), functions that map from patient characteristics to treatment recommendations. Ideally, we would like to learn ITRs that if followed in practice would lead to better outcomes on average in the target population than if another treatment strategy was used, e.g., a one-size-fits-all approach. These ITRs are called optimal ITRs [50, 77].

In the language of reinforcement learning, we focus on the "batch, off-policy" setting. By “batch" we mean that data have been previously collected and no new data will be received, and by “off policy’ we mean that the strategy for assigning treatments in the observed data (e.g., through randomization as in a clinical trial) may not be the optimal strategy (or alternatively, regime or policy) [113]. Within this setting, a large number of methods and approaches have been developed to learn such ITRs from data.

Some approaches estimate the expected value of the outcome we would expect under a particular ITR without any parametric assumptions. Then, the optimal ITR may be learned by searching over a class of ITRs. Examples of this general strategy include those proposed by [83], [100], and [139].

Another class of approaches, sometimes referred to as indirect methods, model the mean of the outcome conditional on treatment and covariates. For a multi-stage ITR, sometimes called a dynamic treatment regime, this entails specifying a conditional mean model for each stage. Through these estimated conditional means, optimal ITRs can be deduced. One of the most popular regression-based frameworks for learning optimal ITRs is Q-learning [126, 75]. Q-learning has been used to learn optimal ITRs in many settings, including clinical trial data [103], observational data [73], and in the presence of censoring for time-to-event data [32].

Machine learning or classification-based optimal ITR learning approaches convert the optimal ITR learning problem into the classification framework by which machine learning methods can be employed. [143] introduced outcome-weighted learning (OWL) which leverages a simple value function estimator and the Radon-Nikodym theorem to rewrite the value function as a weighted classification problem. Consequently, learning the ITR that optimizes the value function can be solved as minimizing the classification loss function. Since its introduction, a number of extensions to OWL have been made including backwards outcome-weighted learning (BOWL) and simultaneous outcome-weighted learning (SOWL) for learning optimal multi-stage treatment regimes [142], residual weighted learning (RWL) [146] and augmented outcome-weighted learning (AOL) [63] which improve the finite sample properties of OWL, robust outcome-weighted learning (ROWL) which uses an angle-based classification approach [26], and efficient augmentation and relaxation learning (EARL) which employs both a propensity model and outcome model and has the double robustness property [141].

Finally, a few Bayesian approaches for learning optimal ITRs have also been proposed. The Bayesian machine learning (BML) approach was introduced by [78]. It employs Bayesian modeling within a framework that closely aligns with Q-learning by modeling the outcomes at each stage. Likelihood-based approaches, strategies that model both the distribution of the final outcome and the intermediate outcomes, have also been proposed within the Bayesian framework [116, 117, 4, 138, 129, 136].

The focus of this paper will be on the machine learning, or classification-based method, for learning optimal ITRs. Although classification-based approaches are powerful and avoid estimating models that are not the target of the analysis itself, there are limitations. For example, many machine learning methods for classification do not naturally quantify uncertainty, e.g., quantification of the uncertainty of a particular prediction. While this may be acceptable in some cases, being unable to quantify uncertainty is a serious gap when generating evidence for health care decision-making. Moreover, machine learning analyses are often evaluated in terms of predictive power which does not necessarily translate into inferential capability.

In this paper, we present a Bayesian approach to OWL. To our knowledge, this is the first Bayesian optimal ITR learning strategy to directly learn optimal ITRs. Using a construction similar to [88], we construct a pseudo-likelihood from the weighted classification loss function. Once transformed from an optimization-based framework to a probabilistic framework, our method generates an entire posterior distribution that can be used for inference and, most powerfully, for uncertainty quantification of the treatment recommendations themselves. Our main contributions are as follows:

1. We propose a Bayesian approach to learning optimal ITRs that leverages the classification-based framework and avoids modeling the outcome or nuisance conditional mean models.
2. We propose a simple Gibbs sampling algorithm for learning such an optimal ITR.
3. We demonstrate how to use our resulting pseudo-posterior distribution to quantify uncertainty in the treatment recommendations.

In Section 5.2, we set the notation and review OWL and Bayesian support vector machines. In Section 5.3 we construct the probabilistic formulation of the OWL classification problem, derive a Gibbs sampling algorithm for estimation, and detail our approach to uncertainty quantification. In Section 5.4 we demonstrate the performance of our approach through simulation studies. We conclude in Section 5.5 with a discussion of our results and future work.

5.2 Background

5.2.1 Setting

We let $A \in \mathcal{A} = \{-1, 1\}$ denote the action, or treatment, and assume that observed treatments are assigned randomly as in a clinical trial with $P(A = 1) = \rho$ known. Let $X_i = (X_{i,1}, \ldots, X_{i,p})^\top \in \mathcal{X}$ denote the p-dimensional biomarker and prognostic information vector, and let R denote the outcome (bigger is better). We further assume that the reward can be rescaled so that $R > 0$. Then, the observed data is iid replicates of (A_i, X_i, R_i) for $i = 1, \ldots, n$.

An ITR is a function d that maps from $\mathcal{X}$ to a recommended treatment in $\mathcal{A}$. For a given ITR d, the value of d is $V(d) = \mathbb{E}[R(d)]$, where $R(d)$ is the reward we would observe if treatments were allocated according to rule d. An optimal ITR d^{opt} satisfies $V(d^{\text{opt}}) \geq V(d)$ for all $d \in \mathcal{D}$, where $\mathcal{D}$ is a class of ITRs. Our goal is to learn an optimal ITR d^{opt}. Under the assumptions of causal consistency, the stable unit treatment value assumption, no unmeasured confounding, and positivity, $V(d)$ can be identified from the observed data and $V(d) = \mathbb{E}\{\max_{A \in \mathcal{A}} \mathbb{E}[R|A = d(\mathbf{x}), X = \mathbf{x}]\}$.

5.2.2 Outcome-weighted learning

If we let P denote the distribution of (X, A, R), and P^d denote the distribution of (X, A, R) when $A = d(X)$, then the reward we would expect if ITR $d(X)$ were followed is given by

$$\mathbb{E}^d(R) = \int R dP^d = \int R \frac{dP^d}{dP} dP = \mathbb{E}\left[\frac{\mathbb{1}(A = d(X))}{A\rho + (1 - A)/2} R\right]. \tag{5.1}$$

[143] showed that maximizing Equation (5.1) is equivalent to a weighted classification problem and thereby solvable using techniques from machine learning. Specifically, they proposed OWL, a strategy for learning an optimal ITR using a convex surrogate

loss function in place of the zero-one loss function and strategies from support vector machines. OWL minimizes the objective function

$$Q_n^{\text{OWL}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{A_i\rho + (1-A_i)/2}(1 - A_i h(X_i, \boldsymbol{\beta}))_+ \tag{5.2}$$

where $(z)_+ = max(z, 0)$ denotes the hinge loss function and $h(\cdot)$ is the ITR parameterized by $\boldsymbol{\beta}$. The article [110] introduced a penalized variant of OWL that included a regularization term $p_\lambda(\boldsymbol{\beta})$ for the ITR parameters. POWL minimizes the objective function

$$Q_n^{\text{POWL}}(\boldsymbol{\beta})\frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{A_i\rho + (1-A_i)/2}(1 - A_i h(X_i, \boldsymbol{\beta}))_+ + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{5.3}$$

where $p_\lambda(\boldsymbol{\beta})$ is a penalty function and λ is a tuning parameter.

5.2.3 Bayesian support vector machines

Although the pure machine learning framework is powerful, it is limited in its ability to capture and model uncertainty as in a statistical framework. The article [88] bridged this gap between pure machine learning and statistical modeling for SVMs by showing how to cast SVM into a Bayesian framework. They considered the L^α-norm regularized support vector classifier that chooses β to minimize

$$d_\alpha(\beta, \nu) = \sum_{i=1}^{n} \max(1 - r_i \mathbf{x}_i^\top \beta, 0) + \nu^{-\alpha} \sum_{j=1}^{k} |\beta_j/\sigma_j|^\alpha \tag{5.4}$$

where σ_j is the standard deviation of the j-th element of $\mathbf{x}$ and ν is a tuning parameter. For this objective function, the learned classifier is a linear classifier. The article [88] shows that minimizing Equation (5.4) is equivalent to finding the mode of the pseudo-posterior distribution $p(\beta|\nu, \alpha, y)$

$$\begin{aligned} p(\beta|\nu, \alpha, r) &\propto \exp(-d_\alpha(\beta, \nu)) \\ &\propto C_\alpha(\nu) L(r|\beta) p(\beta|\nu, \alpha) \end{aligned} \tag{5.5}$$

where C_α is a pseudo-posterior normalization constant. Thus, the data dependent factor $L(y|\beta)$ is a pseudo-likelihood

$$L(r|\beta) = \prod_i L_i(r_i|\beta) = \exp\left\{-2\sum_{i=1}^{n} \max(1 - r_i x_i^\top \beta, 0)\right\}. \tag{5.6}$$

The main theoretical result from [88] is that the pseudo-likelihood contribution $L_i(r_i|\beta)$ is a location-scale mixture of normals ([88], Theorem 1).

5.3 Our approach

We follow the strategy employed by [88] to cast the OWL objective function into a probabilistic Bayesian learning framework. The conversion is not one-to-one since [88] constructed a Bayesian model for a standard SVM whereas the objective function for OWL Equation (5.2) is a weighted SVM problem. We first employ a non-penalty prior to mimic the original formulation of OWL as in [143], and later we demonstrate the inclusion of a penalty prior on the ITR coefficients. Minimizing Equation (5.2) is equivalent to finding the mode of the pseudo-posterior which we can write as

$$\begin{aligned} p(\mathbf{x}|a_i,\nu,\alpha) &\propto \exp(-Q_n(\boldsymbol{\beta},\nu,\alpha)) \\ &\propto \exp\left\{\sum_{n=1}^{n}\frac{r_i}{a_i\rho+(1-A_i)/2}(1-a_i h(\mathbf{x}_i,\boldsymbol{\beta}))_+\right\}\prod_{j=1}^{p}p(\beta_j|\mu_0,\sigma_0^2) \\ &\propto C(\nu,\alpha)L(a|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mu_0,\sigma_0^2). \end{aligned} \tag{5.7}$$

Throughout, we will assume $R > 0$. When this is not the case, a distance-preserving transformation of R from $\mathbb{R}$ to $\mathbb{R}^+$ can be used. Assuming that h is linear, i.e., $h(\mathbf{x}_i,\boldsymbol{\beta}) = \mathbf{x}_i^\top\boldsymbol{\beta}$ and following the strategy taken in Theorem 1 of [88], the contribution of a single observation to the pseudo-likelihood is given by

$$\begin{aligned} L_i(a_i|r_i,\mathbf{x}_i,\boldsymbol{\beta}) =& \exp\left\{-2\frac{r_i}{a_i\rho+(1-a_i)/2}\max(1-a_i\mathbf{x}_i^\top\boldsymbol{\beta},0)\right\} \\ =& \mathbb{1}(a_i=1)\int_0^\infty\frac{1}{\sqrt{2\pi\lambda_i}}\exp\left\{-\frac{1}{2\lambda_i}\left(\frac{r_i}{\rho}+\lambda_i-\frac{r_i}{\rho}a_i\mathbf{x}_i^\top\boldsymbol{\beta}\right)^2\right\}d\lambda_i \\ &+\mathbb{1}(a_i=-1) \\ &\times\int_0^\infty\frac{1}{\sqrt{2\pi\lambda_i}}\exp\left\{-\frac{1}{2\lambda_i}\left(\frac{r_i}{1-\rho}+\lambda_i-\frac{r_i}{1-\rho}a_i\mathbf{x}_i^\top\boldsymbol{\beta}\right)^2\right\}d\lambda_i, \end{aligned} \tag{5.8}$$

or in other words that $L_i(a_i,\lambda_i|r_i,\mathbf{x}_i,\boldsymbol{\beta})$ is a scale mixture of Gaussians.

Proof. Because we have assumed that the reward is is strictly positive, the weight $\frac{r_i}{a_i\rho+(1-a_i)/2}$ is also positive and can be brought inside of the maximization operator so that

$$\begin{aligned} &L_i(a_i|r_i,\mathbf{x}_i,\boldsymbol{\beta}) \\ &\quad=\exp\left\{-2\frac{r_i}{a_i\rho+(1-a_i)/2}\max(1-a_i\mathbf{x}_i^\top\boldsymbol{\beta},0)\right\} \\ &\quad=\exp\left\{-2\max\left(\frac{r_i}{a_i\rho+(1-a_i)/2}(1-a_i\mathbf{x}_i^\top\boldsymbol{\beta},0)\right)\right\} \\ &\quad=\mathbb{1}(a_i=1)\exp\left\{-2\max\left(\frac{r_i}{\rho}(1-a_i\mathbf{x}_i^\top\boldsymbol{\beta},0)\right)\right\} \\ &\qquad+\mathbb{1}(a_i=-1)\exp\left\{-2\max\left(\frac{r_i}{1-\rho}(1-a_i\mathbf{x}_i^\top\boldsymbol{\beta},0)\right)\right\}. \end{aligned}$$

The derivation of the pseudolikelihood representation follows [88]: Andrews and Mallows (1974) showed that $\int_0^\infty \frac{a}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2}(a^2\lambda + b^2\lambda^{-1})} d\lambda = e^{-|ab|}$. Setting $a = 1$ and $b = u$, we have

$$\int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2}(\lambda + u^2\lambda^{-1})} d\lambda = e^{-|u|}.$$

Multiplying through by e^{-u} and recalling the identity $\max(u, 0) = \frac{1}{2}(|u| + u)$, we have

$$\begin{aligned}
& e^{-u} \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2}(\lambda + u^2\lambda^{-1})} d\lambda = e^{-|u|} e^{-u} \\
\implies & \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2}(\lambda + u^2\lambda^{-1}) - u} d\lambda = e^{-|u| - u} \\
\implies & \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2\lambda}(\lambda^2 + u^2 + 2u\lambda)} d\lambda = e^{-|u| - u} \\
\implies & \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2\lambda}(u + \lambda)^2} d\lambda = e^{-2\max(u, 0)}.
\end{aligned}$$

Thus we can write the individual contribution of each observation to the marginal likelihood as

$$\begin{aligned}
& L_i(a_i | \lambda_i, r_i, \mathbf{x}_i, \boldsymbol{\beta}) \\
& = \mathbb{1}(a_i = 1) \exp\left\{ -2 \max\left(\frac{r_i}{\rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}, 0) \right) \right\} \\
& \quad + \mathbb{1}(a_i = -1) \exp\left\{ -2 \max\left(\frac{r_i}{1 - \rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}, 0) \right) \right\} \\
& = \mathbb{1}(a_i = 1) \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{ -\frac{1}{2\lambda_i} \left(\frac{r_i}{\rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda_i \right)^2 \right\} d\lambda_i \\
& \quad + \mathbb{1}(a_i = -1) \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{ -\frac{1}{2\lambda_i} \left(\frac{r_i}{1 - \rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda_i \right)^2 \right\} d\lambda_i \\
& = \mathbb{1}(a_i = 1) \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{ -\frac{1}{2\lambda_i} \left(\frac{r_i}{\rho} + \lambda_i - \frac{r_i}{\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 \right\} d\lambda_i \\
& \quad + \mathbb{1}(a_i = -1) \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{ -\frac{1}{2\lambda_i} \left(\frac{r_i}{1 - \rho} + \lambda_i - \frac{r_i}{1 - \rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 \right\} d\lambda_i,
\end{aligned}$$

and that $L_i(a_i, \lambda_i | r_i, \mathbf{x}_i, \boldsymbol{\beta})$ is a scale mixture of Gaussians. □

5.3.1 Prior specification for the ITR parameters

In their formulation of Bayesian SVM, [88] use the exponential power prior for $\boldsymbol{\beta}$, a prior that can be shown to be equivalent to L1-regularization of the regression

parameters. Regularization of the OWL parameters have been explored as in [110]. In this paper, we first construct our method as an analogy to the original formulation of OWL without penalization. We make this choice because (1) our primary aim is to develop a Bayesian classification-based ITR learning approach, and because (2) L1-regularization does not necessarily yield sparse rules (see the discussion in Section 4.1 of [88]). However, regularization help avoid overfitting, a common problem in machine learning. Thus, we also explore penalty priors for $\boldsymbol{\beta}$, including the exponential power prior distribution and the spike-and-slab prior distribution.

5.3.1.1 Normal prior distribution for $\boldsymbol{\beta}$

We first consider the case with normal distribution priors on the treatment rule parameters $\beta_j \sim N(\mu_0, \sigma_0^2)$ for $j = 1, \ldots, p$ where μ_0 and σ_0^2 are hyperparameters. With the pseudo-likelihood and a suitable prior for the ITR parameters defined, we can write the pseudo-posterior distribution as

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{x}, \mathbf{r}, \mathbf{a}, \alpha, \nu) \propto & \prod_{\{i:a_i=1\}}^{n} \lambda_i^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \sum_{i:a_1=1}^{n} \frac{\left(\frac{r_i}{\rho} + \lambda_i - \frac{r_i}{\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\lambda_i} \right\} \\ & \times \prod_{\{i:a_i=-1\}}^{n} \lambda_i^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \sum_{i:a_i=-1}^{n} \frac{\left(\frac{r_i}{1-\rho} + \lambda_i - \frac{r_i}{1-\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\lambda_i} \right\} \\ & \times \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2} \frac{(\beta_j - \mu_{0,j})^2}{\sigma_0^2} \right\} \end{aligned} \tag{5.9}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top$, $\mathbf{r} = (r_i, \ldots, r_n)^\top$, and $\mathbf{a} = (a_1, \ldots, a_n)^\top$.

5.3.2 Exponential power prior distribution for $\boldsymbol{\beta}$

Rather than use normal priors for the coefficients of the rule, [88] employed an exponential power prior on $\boldsymbol{\beta}$. This prior contains the regularization penalty, and from Theorem 2 of [88], the double exponential prior regularization penalty can be written as

$$p(\beta_j|\nu, \alpha = 1) = \int_0^\infty \phi(\beta_j|0, \nu^2 \omega_j \sigma_j^2) \frac{1}{2} e^{-\frac{\omega_j}{2}} d\omega_j \tag{5.10}$$

where $p(\omega_j|\alpha) \propto \omega_j^{-\frac{3}{2}} St_{\alpha/2}^+(\omega_j^{-1})$ and $St_{\alpha/2}^+$ is the density function of a positive stable random variable of index $\alpha/2$. In particular, when $\alpha = 1$, $p(\omega_j|\alpha) \sim Exponential(2)$ (Corollary 1 of [88]).

Under this prior distribution specification, we can write the pseudo-posterior dis-

tribution as

$$p(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\omega}|\mathbf{x},\mathbf{r},\mathbf{a},\alpha,\nu) \propto \prod_{\{i:a_i=1\}}^{n} \lambda_i^{-1/2} \cdot \exp\left\{-\frac{1}{2}\sum_{i:a_1=1}^{n} \frac{\left(\frac{r_i}{\rho}+\lambda_i-\frac{r_i}{\rho}a_i\mathbf{x}_i^\top\boldsymbol{\beta}\right)^2}{\lambda_i}\right\}$$
$$\times \prod_{\{i:a_i=-1\}}^{n} \lambda_i^{-1/2} \cdot \exp\left\{-\frac{1}{2}\sum_{i:a_i=-1}^{n} \frac{\left(\frac{r_i}{1-\rho}+\lambda_i-\frac{r_i}{1-\rho}a_i\mathbf{x}_i^\top\boldsymbol{\beta}\right)^2}{\lambda_i}\right\}$$
$$\times \prod_{j=1}^{p} \omega_j^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\nu^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\sigma_j^2\omega_j}\right\} \cdot \prod_{j=1}^{p} p(\omega_j|\alpha). \tag{5.11}$$

where $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_p)^\top$.

5.3.3 Spike-and-slab prior distribution for $\boldsymbol{\beta}$

The article [88] also explored the use of a spike-and-slab prior for $\boldsymbol{\beta}$. The spike-and-slab prior is a Bayesian approach used for variable selection. It combines a "spike" component, which is a Dirac delta function at zero, to induce sparsity by shrinking some coefficients exactly to zero, and a "slab" component that allows other coefficients to vary freely [71, 30]. Thus, the spike-and-slab prior on the jth coefficient β_j can be written as

$$p(\beta_j|\gamma_j,\nu^2) = \gamma_j N(0,\nu^2\sigma_j^2) + (1-\gamma_j)\delta_0(\beta_j) \tag{5.12}$$

where $\delta_0(\cdot)$ is the Dirac measure (point mass at 0). The prior on γ_j is given by

$$p(\gamma_j|\pi) = \pi^{\gamma_j}(1-\pi)^{1-\gamma_j}. \tag{5.13}$$

Letting $\odot$ denote elementwise multiplication, i.e., where $(a_1,\ldots,a_n)\odot(b_1,\ldots,b_n) = (a_1b_1,\ldots,a_nb_n)$, the full pseudo-posterior when a spike-and-slab prior distribution is specified for $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} p(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\gamma}|\mathbf{y},\mathbf{X},\pi,\nu) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda}) \prod_{j=1}^{p}\left[p(\beta_j|\gamma_j,\nu^2)p(\gamma_j|\pi)\right] \\ &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\lambda_i}}\exp\left\{\frac{1}{2}\frac{(1+\lambda_i-y_ix_i^\top(\boldsymbol{\gamma}\odot\boldsymbol{\beta}))^2}{\lambda_i}\right\} \\ &\quad\times\prod_{j=1}^{p}\left[(\gamma_jN(0,\nu^2\sigma_j^2)+(1-\gamma_j)\delta_0(\beta_j))\pi^{\gamma_j}(1-\pi)^{1-\gamma_j}\right]. \end{aligned} \tag{5.14}$$

5.3.4 Estimation

To draw from the pseudo-posterior distribution, [88] employed two algorithms, an expectation-minimization (EM) approach and a Gibbs sampling approach. The approach we take is the latter. Although sampling the pseudo-posterior is likely to be more time intensive than estimation via the EM algorithm, the rationale for a fully Bayesian approach is to enable uncertainty quantification (Section 5.3.5).

5.3.4.1 Normal prior distribution for $\boldsymbol{\beta}$

The full pseudo-posterior distribution under normal priors for $\boldsymbol{\beta}$ Equation (5.9) has two unknown parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. To sample these parameters, we derive a Gibbs sampling algorithm, which entails sequentially sampling each parameter conditionally on the most up-to-date values of the other parameters. We give a high level summary of the derivation in this section and full details in Section 5.6. The conditional distribution of $\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, \mathbf{a}_i, \mathbf{r}_i$ (up to a normalizing constant) can be written as:

$$p(\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, a_i, r_i) \times \mathbb{1}(a = -1)\lambda_i^{-1/2} \cdot \exp\left\{-\frac{1}{2}\left(\lambda_i + \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})\lambda_i^{-1}\right)\right\}.$$

From [20], page 479, a random variable has the generalized inverse Gaussian distribution $\mathcal{GIG}(\gamma, \psi, \chi)$ if its density function is $p(x|\gamma, \psi, \chi) = C(\gamma, \psi, \chi)x^{\gamma-1}\exp\left\{-\frac{1}{2}\left(\frac{\chi}{x} + \psi x\right)\right\}$, where $C(\gamma, \psi, \chi)$ is a normalization constant. Thus

$$\begin{aligned} p(\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, a_i, r_i) \sim\ & \mathbb{1}(a_i = 1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right) \\ & + \mathbb{1}(a_i = -1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right). \end{aligned} \tag{5.15}$$

The conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ follows from standard arguments for Bayesian linear models. The notable difference from such a standard model is that $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ is a mixture over two distributions, one for when the observed treatment in the data under analysis is 1 and one for when the observed treatment is -1. The conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ has the form

$$\begin{aligned} & p(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{r}, \mathbf{a}, \mathbf{x}) \\ & \propto \exp\left\{-\frac{1}{2}\sum_{\{i:a_i=1\}}\left(-2\frac{r_i}{\rho}a_i\mathbf{x}_i^\top\boldsymbol{\beta}\left(1 + \frac{r_i}{\rho\lambda_i}\right) + \left(\frac{r_i}{\rho}\right)^2\frac{1}{\lambda_i}(a_i\mathbf{x}_i^\top\boldsymbol{\beta})^2\right)\right\} \\ & \quad \cdot \exp\left\{-\frac{1}{2}\sum_{\{i:a_i=-1\}}\left(-2\frac{r_i}{1-\rho}a_i\mathbf{x}_i^\top\boldsymbol{\beta}\left(1 + \frac{1}{(1-\rho)\lambda_i}\right) + \left(\frac{r_i}{1-\rho}\right)^2\frac{1}{\lambda_i}(a_i\mathbf{x}_i^\top\boldsymbol{\beta})^2\right)\right\} \\ & \quad \cdot \exp\left\{-\frac{1}{2}\sum_{j=1}^{p}\frac{(\beta_j - \mu_{0,1})^2}{\sigma_0^2}\right\} \end{aligned}$$

Let $n_1 = \sum_{i=1}^n \mathbb{1}(a_i = 1)$ and $n_{-1} = \sum_{i=1}^n \mathbb{1}(a_i = -1)$. Define $\mathbf{X}_1$, $\mathbf{W}_1$, $\mathbf{R}_1$, and $\boldsymbol{\Lambda}_1$ as

$$\mathbf{X}_1 \equiv \begin{pmatrix} a_1 x_{1,1} & \cdots & a_1 x_{1,p} \\ \vdots & & \vdots \\ a_{n_1} x_{n_1,1} & \cdots & a_{n_1} x_{n_1,p} \end{pmatrix}_{(n_1 \times p)}, \qquad \mathbf{W}_1 \equiv \begin{pmatrix} 1 + \frac{r_1}{\lambda_1} \\ \vdots \\ 1 + \frac{r_{n_1}}{\lambda_{n_1}} \end{pmatrix}_{(n_1 \times 1)},$$

$$\mathbf{R}_1 \equiv diag(r_1/\rho, \ldots, r_{n_1/\rho})_{(n_1 \times n_1)}, \text{ and} \qquad \boldsymbol{\Lambda}_1 = diag(\lambda_1, \ldots, \lambda_{n_1}). \tag{5.16}$$

Define $\mathbf{X}_{-1}$, $\mathbf{W}_{-1}$, $\mathbf{R}_{-1}$, and $\mathbf{\Lambda}_{-1}$ analogously. Additionally define and $\mathbf{\Sigma}$ as $diag(\sigma_1, \ldots, \sigma_p)$.

Then, we have that

$$
\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}) \\
&\propto \exp\left\{ -\frac{1}{2}\Bigg[\boldsymbol{\beta}^\top \underbrace{\left(\mathbf{X}_1^\top \mathbf{R}_1^\top \mathbf{\Lambda}_1^{-1} \mathbf{R}_1 \mathbf{X}_1 + \mathbf{X}_{-1}^\top \mathbf{R}_{-1}^\top \mathbf{\Lambda}_{-1}^{-1} \mathbf{R}_{-1} \mathbf{X}_{-1} + \mathbf{\Sigma}^{-1}\right)}_{\equiv B_1^{-1}} \boldsymbol{\beta} \right. \\
&\quad \left. - 2(\underbrace{\mathbf{W}_1^\top \mathbf{R}_1 \mathbf{X}_1 + \mathbf{W}_{-1}^\top \mathbf{R}_{-1} \mathbf{X}_{-1} + \boldsymbol{\mu}_0^\top \mathbf{\Sigma}^{-1}}_{\equiv b_1})\boldsymbol{\beta}\Bigg]\right\} \\
&\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta} - B_1 b_1)^\top B_1^{-1} (\boldsymbol{\beta} - B_1 b_1) \right\}.
\end{aligned}
$$

In other words, the conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$, $\boldsymbol{\omega}$, and is multivariate normal with mean $B_1 b_1$ and variance-covariance matrix B_1. With the necessarily conditional distributions derived, the Gibbs sampling algorithm for sampling from the posterior distribution is given in Box 1.

Box 1. Gibbs sampling algorithm for normal distribution priors on $\boldsymbol{\beta}$

Initialize $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$; set the hyperparameters $\boldsymbol{\mu}_0$ and σ_0^2.
Step 1: Draw $\boldsymbol{\beta}^{(g+1)}|\boldsymbol{\lambda}^{(g)}, \mathbf{r}, \mathbf{a}, \mathbf{x} \sim \mathcal{N}(B_1^{(g)} b_1^{(g)}, B_1^{(g)})$.
Step 2: Draw $\boldsymbol{\lambda}^{-1(g+1)}|\boldsymbol{\beta}^{(g)}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ where

$$
\begin{aligned}
\lambda_i \sim\ & \mathbb{1}(a_i = 1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right) \\
&+ \mathbb{1}(a_i = -1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right).
\end{aligned}
$$

Repeat Steps 1 and 2 until the chains converge.

5.3.4.2 Exponential power prior distribution for $\boldsymbol{\beta}$

The full pseudo-posterior distribution when an exponential power prior distribution is specified for $\boldsymbol{\beta}$ (Equation (5.11)) has three unknown parameters, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\omega}$. To sample these parameters, we derive a Gibbs sampling algorithm, which entails sequentially sampling each parameter conditionally on the most up-to-date values of the other parameters. We give a high-level summary of the derivation in this section and full details in Section 5.6.3. The conditional distribution of $\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, \mathbf{a}_i, \mathbf{r}_i$ is the same as in the case with normal prior distributions for $\boldsymbol{\beta}$ given in Equation (5.15).

The conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ follows from standard arguments for Bayesian linear models. The notable difference from such a standard model is that $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ is a mixture over two distributions, one for when the observed treatment in the data under analysis is 1 and one for when the observed treatment is -1. The

third term of the conditional distribution is the penalty. The conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ has the form

$$
\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}) \\
&\quad \propto \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=1\}} \left(-2\frac{r_i}{\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \left(1 + \frac{r_i}{\rho\lambda_i}\right) + \left(\frac{r_i}{\rho}\right)^2 \frac{1}{\lambda_i} (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right) \right\} \\
&\quad \cdot \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=-1\}} \left(-2\frac{r_i}{1-\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \left(1 + \frac{1}{(1-\rho)\lambda_i}\right) + \left(\frac{r_i}{1-\rho}\right)^2 \frac{1}{\lambda_i} (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right) \right\} \\
&\quad \cdot \exp\left\{ -\frac{1}{2\nu^2} \sum_{j=1}^{p} \frac{\beta_j^2}{\sigma_j^2 \omega_j} \right\}.
\end{aligned}
$$

Letting $\boldsymbol{\Omega} \equiv diag(\omega_1, \ldots, \omega_p)_{(p\times p)}$, we have that

$$
\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}) \\
&\quad \propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}^\top \underbrace{\left(\mathbf{X}_1^\top \mathbf{R}_1^\top \boldsymbol{\Lambda}_1^{-1} \mathbf{R}_1 \mathbf{X}_1 + \mathbf{X}_{-1}^\top \mathbf{R}_{-1}^\top \boldsymbol{\Lambda}_{-1}^{-1} \mathbf{R}_{-1} \mathbf{X}_{-1} + \nu^{-2} \boldsymbol{\Omega}^{-1} \boldsymbol{\Sigma}^{-1} \right)}_{\equiv B_2^{-1}} \boldsymbol{\beta} \right. \right. \\
&\qquad \left. \left. - 2(\underbrace{\mathbf{W}_1^\top \mathbf{R}_1 \mathbf{X}_1 + \mathbf{W}_{-1}^\top \mathbf{R}_{-1} \mathbf{X}_{-1}}_{\equiv b_2}) \boldsymbol{\beta} \right] \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - B_2 b_2)^\top B_2^{-1} (\boldsymbol{\beta} - B_2 b_2) \right\}.
\end{aligned}
$$

Thus, the conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$, $\boldsymbol{\omega}$, and is multivariate normal with mean $B_2 b_2$ and variance-covariance matrix B_2. Full details of this derivation are given in Section 5.6.

Finally, the conditional distribution of $\boldsymbol{\omega}|\boldsymbol{\beta}, \nu$ is the same as that given in Corollary 3 of [88]. For $\alpha = 1$, the full conditional distribution of ω is $\omega_j^{-1}|\beta_j, \nu \sim \mathcal{IG}(\nu\sigma_j/|\beta_j|, 1)$. Together, with these three conditional distributions, we can summarize the Gibbs sampling algorithm as in Box 2.

Box 2. Gibbs sampling algorithm for exponential power distribution prior on $\boldsymbol{\beta}$

Initialize $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$.
Step 1: Draw $\boldsymbol{\beta}^{(g+1)}|\nu, \boldsymbol{\Lambda}^{(g)}, \boldsymbol{\Omega}^{(g)}, \mathbf{r}, \mathbf{a}, \mathbf{x} \sim \mathcal{N}(B_2^{(g)} b_2^{(g)}, B_2^{(g)})$.
Step 2: Draw $\boldsymbol{\lambda}^{-1(g+1)}|\boldsymbol{\beta}^{(g+1)}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ where

$$\lambda_i^{(g+1)}|\boldsymbol{\beta}^{(g+1)}, \nu, r_i, x_i \sim \mathbb{1}(a_i = 1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}^{(g+1)})^2\right) + \mathbb{1}(a_i = -1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}^{(g+1)})^2\right).$$

Step 3: Draw $\omega_j^{-1(g+1)}|\beta_j^{(g+1)}, \nu \sim \mathcal{IG}(\nu\sigma_j|\beta_j|^{-1}, 1)$
Repeat Steps 1, 2, and 3 until the chains converge.

5.3.4.3 Spike-and-slab prior distribution for $\boldsymbol{\beta}$

The full pseudo-posterior distribution when a spike-and-slab prior is specified for $\boldsymbol{\beta}$ has three unknown parameters, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\gamma}$. Similarly to the procedure outlined earlier, we employ a Gibbs sampling algorithm to obtain these parameters. The conditional distribution of $\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, \mathbf{a}_i, \mathbf{r}_i$ is the same as in the case with normal prior distributions for $\boldsymbol{\beta}$ given in Equation (5.15).

The conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$, $\mathbf{r}$, $\mathbf{a}$, and $\mathbf{x}$ mirrors the previous exponential power prior distribution, but without the third term of the penalty term. The adjustments are slight, with only the following modifications:

$$\begin{aligned} B_\gamma^{-1} &= X_1^T R_1^T \Lambda_1^{-1} R_1 X_1 + X_{-1}^T R_{-1}^T \Lambda_{-1}^{-1} R_{-1} X_{-1} \\ b_\gamma &= B_\gamma (W_1^T R_1 X_1 + W_{-1}^T R_{-1} X_{-1}) \end{aligned}$$

The spike-and-slab prior induces sparsity in coefficients through the parameters γ. The conditional distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{\lambda}$, $\mathbf{r}$, $\mathbf{a}$, and $\mathbf{x}$ may be written as in [88] with b_γ and B_γ previously introduced :

$$p(\gamma|\boldsymbol{\gamma}, \mathbf{r}, \mathbf{a}, \mathbf{x}, \nu) \propto p(\gamma) \frac{|\Sigma_\gamma^{-1}/\nu^2|^{1/2}}{|B_\gamma^{-1}|^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(1 + \lambda_i - a_i x_{i,\gamma}^T b_\gamma)^2}{\lambda_i} - \frac{1}{2\nu^2} b_\gamma^T \Sigma_\gamma^{-1} b_\gamma\right)$$

By exploiting the quadratic term, the conditional distribution can be rewritten as in step 2 of Box 3. This form includes static terms that do not need to be recomputed in every iteration. To sample $\boldsymbol{\gamma}$, a second Gibbs sampler nested within the first must be implemented.

Box 3. Gibbs sampling algorithm for spike-and-slab prior distribution on $\boldsymbol{\beta}$

Initialize $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.
Step 1: Draw $\boldsymbol{\lambda}^{-1(g+1)}|\boldsymbol{\beta}^{(g+1)}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ where

$$\lambda_i^{(g+1)}|\boldsymbol{\beta}^{(g+1)}, \nu, r_i, x_i \sim \mathbb{1}(a_i = 1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}^{(g+1)})^2\right) + \mathbb{1}(a_i = -1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}^{(g+1)})^2\right).$$

Step 2: For $i = 1, \ldots, k$ draw γ_i from $p(\gamma_i|\gamma_{-i})$ which is proportional to

$$p(\gamma|\boldsymbol{\gamma}, \mathbf{r}, \mathbf{a}, \mathbf{x}, \nu) \propto p(\gamma)\frac{|\Sigma_\gamma^{-1}/\nu^2|^{1/2}}{|B_\gamma^{-1}|^{1/2}} \times \exp(-\frac{1}{2}[c(\lambda) + b_\gamma^T (X^T \Lambda^{-1} X)_\gamma b_\gamma - 2b_\gamma^T [X^T(1 + \lambda^{-1})]_\gamma] - \frac{1}{2\nu^2} b_\gamma^T \Sigma_\gamma^{-1} b_\gamma)$$

Step 3: When $\gamma_i = 1$, draw $\boldsymbol{\beta}_\gamma^{(g+1)}|\nu, \boldsymbol{\Lambda}^{(g)}, \mathbf{r}, \mathbf{a}, \mathbf{x} \sim \mathcal{N}(b_\gamma^{(g)}, B_\gamma^{(g)})$.
Repeat Steps 1, 2, and 3 until the chains converge.

5.3.5 Prediction and uncertainty quantification

Using the posterior predictive distribution, we can make treatment recommendations for a new patient and quantify our uncertainty in our recommendation. Let $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\lambda}\}$ and $\tilde{a}$ denote the recommended treatment for a new patient with features $\tilde{\mathbf{x}}$. Then

$$p(\tilde{a} = 1|\tilde{\mathbf{x}}, \mathbf{X}, \mathbf{r}, \mathbf{a}) = \int_\Theta p(\tilde{a} = 1|\tilde{\mathbf{x}}, \mathbf{X}, \mathbf{r}, \mathbf{a}, \boldsymbol{\beta}, \boldsymbol{\lambda}) p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{x}, \mathbf{r}, \mathbf{a}) d\theta.$$

For the class predictions, we can use the probit model which has the form

$$p(a = 1|x) = \Phi(x^\top \beta)$$

where Φ is the cumulative distribution function of the standard normal distribution. Thus we can write the posterior predictive distribution as

$$p(\tilde{a} = 1|\tilde{\mathbf{x}}, \mathbf{X}, \mathbf{r}, \mathbf{a}) = \int_\Theta \Phi(\tilde{\mathbf{x}}^\top \boldsymbol{\beta}) p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{x}, \mathbf{r}, \mathbf{a}) d\theta. \tag{5.17}$$

5.4 Simulation studies

5.4.1 Classification performance

We conducted simulation studies to assess the classification performance of the proposed method, following [143] and [110]. We compared the performance of OWL,

Bayesian OWL with normal priors for $\boldsymbol{\beta}$, Bayesian OWL with exponential power prior for $\boldsymbol{\beta}$, and Bayesian OWL with spike-and-slab prior for $\boldsymbol{\beta}$. For each simulated patient, we generated a 10-dimensional vector of patient features, $X_1, \ldots, X_{10}$, drawn independently and uniformly distributed on $[-1, 1]$. Treatment A was drawn from $\{-1, 1\}$ independently of the prognostic variables with $\mathbb{P}(A = 1) = 1/2$. The outcome variable R was normally distributed with mean $Q_0 = 1 + 2X_1 + X_2 + 0.5X_3 + T_0(X, A)$ and standard deviation 1, where $T_0(X, A)$ was the interaction term between treatment and patient features. We examined two scenarios for the treatment-feature interaction term:

— Scenario 1 : $T_0(A, X) = (X_1 + X_2)A$

— Scenario 2 : $T_0(A, X) = 0.442(1 - X_1 - X_2)A$

Both scenarios 1 and 2 had linear decision boundaries determined by X_1 and X_2. For scenario 1, the true optimal rule was given by $\mathbb{1}(X_1 + X_2 > 0)$, while for scenario 2, it was $\mathbb{1}(1 - X_1 - X_2 > 0)$. OWL was implemented with a linear kernel. For Bayesian OWL, Gibbs sampling was used to draw from the posterior distributions of the parameters 500 times. The first 150 draws were discarded as "burn-in" and point estimates of $\boldsymbol{\beta}$ were computed by taking the mean of the draws from the posterior distribution. Throughout, we set the hyperparameter $\nu = 0.8$.

For each scenario, we varied the training dataset from 100 to 200, 400 and 800 and tested on 1000 patients. For each training set size, we conducted 200 simulation runs. We evaluated classification performance using the misclassification rate, the ratio of the number of patients recommended a treatment counter to the true optimal rule divided by the total number of patients in the simulation run ($\frac{\text{Number of patients misclassified}}{\text{Total number of patients}}$). The simulation results are presented in Table 1 and 2.

n	OWL	Bayesian OWL Normal Prior	Bayesian OWL Exponential Power Prior	Bayesian OWL Spike and Slab
100	0.24	0.38	0.38	0.39
200	0.18	0.34	0.34	0.34
400	0.13	0.29	0.29	0.30
800	0.10	0.24	0.24	0.26

Table 5.1 – Misclassification rates for different methods and sample sizes for scenario 1.

As expected, the classification performance improved among all the ITR learning methods evaluated as the sample size increased. However, OWL consistently outperformed Bayesian OWL in all sample sizes and in both scenarios. We hypothesize that, with additional hyperparameter tuning, the performance of Bayesian OWL can be improved. Ordinarily, one would be hesitant to propose a method that is dominated by an existing method. However, the dominance of OWL is with respect to the misclassification rate. OWL, even with 800 samples in our simulation, has a 10% misclassification rate, and there is no way to determine which of the 10% of the simulated patients are likely misclassified (given a non-optimal treatment recommendation). In contrast,

n	OWL	Bayesian OWL Normal Prior	Bayesian OWL Exponential Power Prior	Bayesian OWL Spike and Slab
100	0.22	0.38	0.38	0.39
200	0.15	0.34	0.34	0.34
400	0.13	0.31	0.31	0.30
800	0.10	0.25	0.25	0.22

Table 5.2 – Misclassification rates for different methods and sample sizes for scenario 2.

Bayesian OWL yields the entire posterior distribution of the estimated optimal ITR and thereby allows for immediate uncertainty quantification of individual-level treatment recommendations. In essence, Bayesian OWL can inform us of which treatment recommendations it is less certain about whereas OWL cannot. We demonstrate this in Section 5.4.2.

5.4.2 Treatment recommendation uncertainty quantification

To highlight the utility of quantifying the uncertainty of individual-level treatment recommendations, we generated a data set of 1000 patients under Scenario 1. We used these simulated data to train a Bayesian OWL model using the exponential power prior. Next, we used the same generative approach to simulate another 1000 patients. Specifically, we used a fine grid to generate X_1 and X_2, the key variables in the true optimal rule (i.e., tailoring variables). The rationale for this approach was to generate a simulated set of patients whose characteristics covered the domain of the true optimal ITR so that we could estimate the uncertainty for combinations of X_1 and X_2 throughout the domain, $\mathcal{X}_1 \times \mathcal{X}_2 \in [-1, 1]^2$.

We evaluated the coefficients for the trained Bayesian OWL model. With an exponential power prior on $\boldsymbol{\beta}$, which is analogous to $L1$ regularization, we would expect the coefficients of the features in the true optimal rule (tailoring variables) to be large and the coefficients of the features not in the rule to be driven close to 0. The magnitudes of the coefficients are displayed in Figure 5.1. As we would hope based on our knowledge of the true optimal rule, the magnitudes of the coefficients for X_1 and X_2 were larger than those for the other features, indicating that the estimated ITR using Bayesian OWL made decisions based on the correct patient features.

Figure 5.2 demonstrates Bayesian OWL's ability to quantify uncertainty in its treatment recommedations. In Scenario 1, the true optimal ITR divides patients into two groups: patients in the upper-right half of the graph (where $X_1 + X_2 > 0$) should ideally get treatment $A = 1$, while those in the lower-left half should get treatment $A = -1$. Using the posterior predictive distribution as in Section 5.3.5, we computed the uncertainty associated with recommending Treatment 1 for patients whose features X_1 and X_2 lie in the upper-right half of the graph and the uncertainty associated with recommending Treatment -1 in the lower-left half. Notably, uncertainty was evaluated

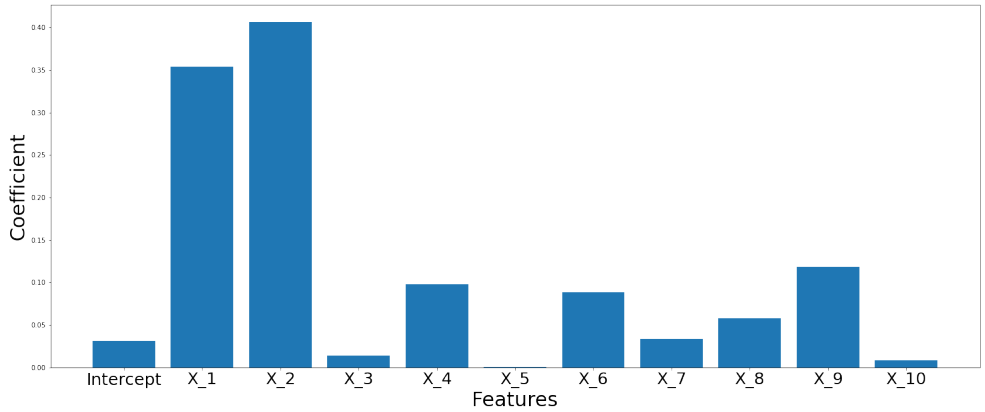

Figure 5.1 – Feature importance

individually for each of the 1000 patients in our test cohort. This means that for each individual, we estimated how certain or uncertain we were about their specific treatment recommendation given their features X_1 and X_2. Moreover, because our test set included simulated patients whose features spanned the domain of the ITR, we were able to compute uncertainty for every "type" of patient who could be recommended a treatment using the estimated ITR. We visualize the uncertainty across the domain of the ITR using a heat map (Figure 5.2). Certainties close to 1 (less uncertainty) are lighter in color and depicted with yellow and light green. In contrast, certainties close to 0 (more uncertain) are darker in color and depicted with purple and dark blue. As expected, Bayesian OWL is more certain about treatment recommendations for patients whose features are far from the decision boundary than for those that are close to the decision boundary.

Furthermore, in Figure 5.2, we have included misclassified individuals in our simulation. Those who were recommended treatment -1 but should have been recommended treatment 1 are indicated by red points, and those who were recommended treatment 1 but should have been recommended treatment -1 are indicated by orange points. We observe that the misclassified patients are located near the boundary, as we expect, and most noteworthy that they are located in regains where the model exhibits the greatest uncertainty (regions in which the background is shaded purple).

5.5 Discussion

In this paper, we introduced a Bayesian formulation of OWL. To our knowledge, this is the first Bayesian strategy for *directly* learning an ITR. Moreover, we demonstrate how the Bayesian approach can enable us to quantify the uncertainty in our treatment recommendations. Both tasks, learning the optimal ITR and uncertainty quantification, can be implemented through a simple Gibbs sampling strategy for sampling the posterior distribution.

One may wonder why a Bayesian approach to OWL is needed since we already have

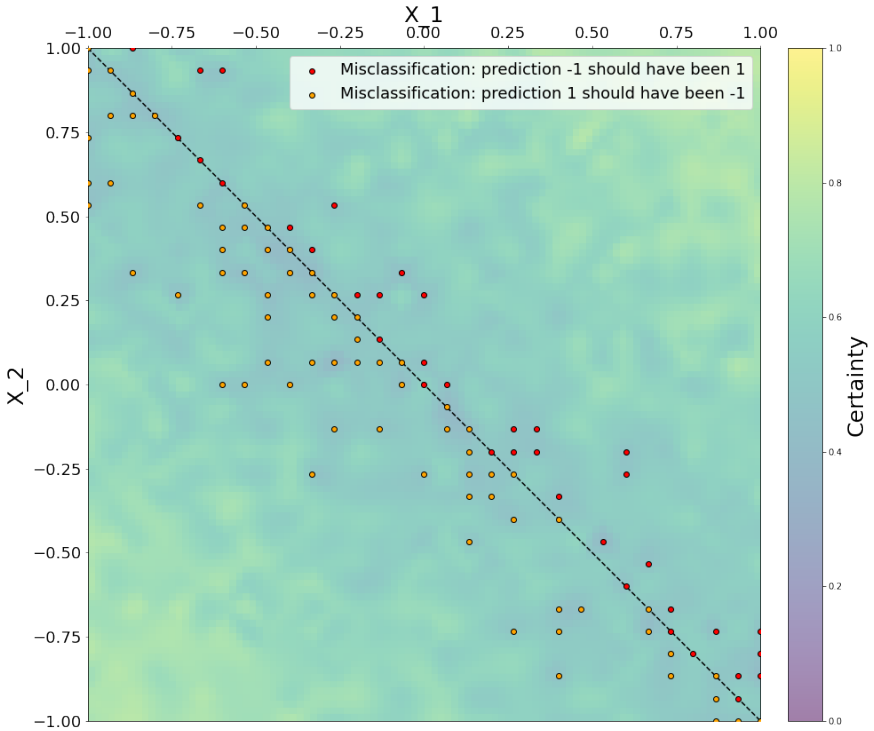

Figure 5.2 – Heatmap of uncertainty quantification

OWL which can rely on methods from convex optimization. We believe that the answer to this question is two-fold. First, OWL is an appealing strategy for learning optimal ITRs because it models the decision rule directly rather than modeling conditional mean models and backing out the optimal ITR. Second, our Bayesian formulation of OWL yields a full pseudo-posterior distribution which means that we can quantify uncertainty in ITR's treatment recommendations *at the individual level*. This is in contrast to the more common approach in the ITR literature which involves estimating the uncertainty in the value of the proposed ITR, i.e., the uncertainty in the expected value we would observe if everyone in the population were treated according to the rule. This may prove useful in the design and implementation of clinical studies by providing a strategy for identifying the types of patients for whom we feel confident in our ability to make treatment recommendations and the patient types that may require additional sampling and information to improve the recommendations. By casting the direct ITR learning approach into a probabilistic framework, we have widened the inferential possibilities for directly learned ITRs.

Our work has limitations. For example, we only examine linear rules. Although linear rules are simple to understand, there may be times when a nonlinear rule is desired or a nonlinear rule significantly outperforms a linear rule, i.e., clinically meaningful improvement from the nonlinear rule is worth the decrease in interpretability. [36] proposed a strategy for Bayesian SVM for nonlinear decision boundaries, which is likely a good blueprint for extending this work to the nonlinear rule setting. Moreover,

our approach does not attempt to do variable selection. This limits its applicability in high dimensional settings in which there is no information or weak information as to which tailoring covariates should be included in the treatment rule. [88] considers L1-regularization on the decision boundary coefficients as well as a spike-and-slab prior to induce sparsity. These may be reasonable strategies to approximate the penalized version of OWL introduced by [110]. Finally, in simulation studies, OWL outperforms Bayesian OWL with respect to the misclassification rate, which is only ameliorated by the fact that Bayesian OWL can tell us when it is uncertain about its recommendations whereas OWL cannot.

While the direct learning, or classification, approach to learning optimal ITRs is incredibly powerful, leveraging tools from machine learning, inference and uncertainty quantification at the individual treatment recommendation level continues to be a challenge. Bayesian OWL overcomes this limitation by fully leveraging the benefits of direct-learning and the use of a probabilistic framework. Generating precision medicine evidence with wider inferential potential can improve our ability to build trust in these treatment algorithms and ultimately improve how we deploy precision medicine evidence in real-world health care decision making.

5.6 Appendix : derivation of the Gibbs sampling algorithms

5.6.1 Conditional distribution of $\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, \mathbf{a}_i, \mathbf{r}_i$

$$
\begin{aligned}
&p(\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, a_i, r_i) \\
&\quad \propto \mathbb{1}(a=1)\lambda_i^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \frac{\left(\lambda_i - \frac{r_i}{\rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})\right)^2}{\lambda_i} \right\} \\
&\quad \times \mathbb{1}(a=-1)\lambda_i^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \frac{\left(\lambda_i - \frac{r_i}{1-\rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})\right)^2}{\lambda_i} \right\} \\
&\quad = \mathbb{1}(a=1)\lambda_i^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \frac{\left(\lambda_i^2 - 2\lambda_i \left(\frac{r_i}{\rho}\right)(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}) + \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right)}{\lambda_i} \right\} \\
&\quad \times \mathbb{1}(a=-1)\lambda_i^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \frac{\left(\lambda_i^2 - 2\lambda_i \left(\frac{r_i}{1-\rho}\right)(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}) + \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right)}{\lambda_i} \right\} \\
&\quad \propto \mathbb{1}(a=1)\lambda_i^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \left(\lambda_i + \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})\lambda_i^{-1}\right) \right\} \\
&\quad \times \mathbb{1}(a=-1)\lambda_i^{-1/2} \cdot \exp\left\{ -\frac{1}{2} \left(\lambda_i + \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})\lambda_i^{-1}\right) \right\}
\end{aligned}
$$

From [20], page 479, a random variable has the generalized inverse Gaussian distribution $\mathcal{GIG}(\gamma, \psi, \chi)$ if its density function is

$$
p(x|\gamma, \psi, \chi) = C(\gamma, \psi, \chi) x^{\gamma-1} \exp\left\{ -\frac{1}{2}\left(\frac{\chi}{x} + \psi x\right)\right\},
$$

where $C(\gamma, \psi, \chi)$ is a normalization constant. Thus

$$
\begin{aligned}
&p(\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, a_i, r_i) \\
&\quad \sim \mathbb{1}(a_i=1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right) \\
&\quad + \mathbb{1}(a_i=-1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right).
\end{aligned}
$$

Recall that if $X \sim \mathcal{GIG}\left(\frac{1}{2}, \lambda, \chi\right)$, then $X^{-1} \sim \mathcal{IG}(\mu, \lambda)$ where $\chi = \lambda/\mu^2$ and $\mathcal{IG}$ denotes the inverse Gaussian distribution. Consequently, we can write

$$
\begin{aligned}
&p(\lambda_i|\boldsymbol{\beta}, \mathbf{x}_i, a_i, r_i) \\
&\quad \sim \mathbb{1}(a_i = 1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right) \\
&\quad + \mathbb{1}(a_i = -1)\mathcal{GIG}\left(\frac{1}{2}, 1, \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2\right).
\end{aligned}
$$

5.6.2 Conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\mu}_0, \sigma_0^2, \mathbf{r}, \mathbf{a}, \mathbf{x}$ (Normal prior)

$$
\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\mu}_0, \sigma_0^2, \mathbf{r}, \mathbf{a}, \mathbf{x}) \\
&\quad \propto \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=1\}} \frac{\left(\frac{r_i}{\rho} + \lambda_i - \left(\frac{r_i}{\rho}\right) a_i \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\lambda_i} \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=-1\}} \frac{\left(\frac{r_i}{1-\rho} + \lambda_i - \left(\frac{r_i}{1-\rho}\right) a_i \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\lambda_i} \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{j=1}^{p} \frac{(\beta_j - \mu_{0,j})^2}{\sigma_0^2} \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=1\}} \frac{\left(\lambda_i + \frac{r_i}{\rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})\right)^2}{\lambda_i} \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=-1\}} \frac{\left(\lambda_i + \frac{r_i}{1-\rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})\right)^2}{\lambda_i} \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{j=1}^{p} \frac{\beta_j^2 - 2\beta_j \mu_{0,j} + \mu_{0,j}^2}{\sigma_0^2} \right\} \\
&\quad = \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=1\}} \frac{\lambda_i^2 + 2\lambda_i \frac{r_i}{\rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}) + \left(\frac{r_i}{\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2)}{\lambda_i} \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=-1\}} \frac{\lambda_i^2 + 2\lambda_i \frac{r_i}{1-\rho}(1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta}) + \left(\frac{r_i}{1-\rho}\right)^2 (1 - a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\lambda_i} \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{j=1}^{p} \frac{\beta_j^2 - 2\beta_j \mu_{0,j} + \mu_{0,j}^2}{\sigma_0^2} \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=1\}} \left(-2\frac{r_i}{\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} + \left(\frac{r_i}{\rho}\right)^2 \frac{1}{\lambda_i} (-2a_i \mathbf{x}_i^\top \boldsymbol{\beta} + (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2) \right) \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=-1\}} \left(-2\frac{r_i}{1-\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} + \left(\frac{r_i}{1-\rho}\right)^2 \frac{1}{\lambda_i} (-2a_i \mathbf{x}_i^\top \boldsymbol{\beta} + (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2) \right) \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{j=1}^{p} \frac{\beta_j^2 - 2\beta_j \mu_{0,j}}{\sigma_0^2} \right\}
\end{aligned}
$$

$$
\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\mu}_0, \sigma_0^2, \mathbf{r}, \mathbf{a}, \mathbf{x}) \\
&\quad \propto \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=1\}} \left(-2\frac{r_i}{\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \left(1 + \frac{r_i}{\rho\lambda_i}\right) + \left(\frac{r_i}{\rho}\right)^2 \frac{1}{\lambda_i} (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right) \right\} \\
&\quad \cdot \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=-1\}} \left(-2\frac{r_i}{1-\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \left(1 + \frac{1}{(1-\rho)\lambda_i}\right) + \left(\frac{r_i}{1-\rho}\right)^2 \frac{1}{\lambda_i} (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right) \right\} \\
&\quad \cdot \exp\left\{ -\frac{1}{2} \sum_{j=1}^{p} \frac{\beta_j^2 - 2\beta_j \mu_{0,j}}{\sigma_0^2} \right\}
\end{aligned}
$$

Consider the summation inside the exponential of the first term, we have

$$
\sum_{\{i:a_i=1\}} \left(-2 \underbrace{\frac{r_i}{\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \left(1 + \frac{r_i}{\rho\lambda_i}\right)}_{\text{Term 1}} + \underbrace{\left(\frac{r_i}{\rho}\right)^2 \frac{1}{\lambda_i} (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2}_{\text{Term 2}} \right).
$$

Let $n_1 = \sum_{i=1}^{n} \mathbb{1}(a_i = 1)$ Working with the summation over the first term

$$
\begin{aligned}
\text{Term 1} = &\sum_{\{a_i=1\}} \frac{r_i}{\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \left(1 + \frac{r_i}{\rho\lambda_i}\right) \\
= &\frac{r_1}{\rho} a_1 \mathbf{x}_1^\top \boldsymbol{\beta} \left(1 + \frac{r_1}{\rho}\lambda_1\right) + \cdots + \frac{r_{n_1}}{\rho} a_{n_1} \mathbf{x}_{n_1}^\top \boldsymbol{\beta} \left(1 + \frac{r_{n_1}}{\rho}\lambda_{n_1}\right) \\
= &\left(\tfrac{r_1}{\rho} \left(1 + \tfrac{r_1}{\rho\lambda_1}\right), \cdots, \tfrac{r_{n_1}}{\rho} \left(1 + \tfrac{r_{n_1}}{\rho\lambda_{n_1}}\right) \right) \begin{pmatrix} a_1 \mathbf{x}_1^\top \boldsymbol{\beta} \\ \vdots \\ a_{n_1} \mathbf{x}_{n_1}^\top \boldsymbol{\beta} \end{pmatrix}
\end{aligned}
$$

Define $\mathbf{X}_1$, $\mathbf{W}_1$, and $\mathbf{R}_1$ as

$$\mathbf{X}_1 \equiv \begin{pmatrix} a_1 x_{1,1} & \cdots & a_1 x_{1,p} \\ \vdots & & \vdots \\ a_{n_1} x_{n_1,1} & \cdots & a_{n_1} x_{n_1,p} \end{pmatrix}_{(n_1 \times p)}, \qquad \mathbf{W}_1 \equiv \begin{pmatrix} 1 + \frac{r_1}{\lambda_1} \\ \vdots \\ 1 + \frac{r_{n_1}}{\lambda_{n_1}} \end{pmatrix}_{(n_1 \times 1)}, \text{ and}$$

$$\mathbf{R}_1 \equiv diag(r_1/\rho, \ldots, r_{n_1/\rho})_{(n_1 \times n_1)}.$$

Then,

$$\text{Term 1} = \frac{1}{\rho} \mathbf{W}_1^\top \mathbf{R}_1 \mathbf{X}_1 \boldsymbol{\beta}$$

because

$$\begin{aligned} &\mathbf{W}_1^\top \mathbf{R}_1 \mathbf{X}_1 \boldsymbol{\beta} \\ &= \begin{pmatrix} 1 + \frac{r_1}{\lambda_1} & \cdots & 1 + \frac{r_{n_1}}{\lambda_{n_1}} \end{pmatrix} \begin{pmatrix} r_1/\rho & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_{n_1}/\rho \end{pmatrix} \begin{pmatrix} a_1 x_{1,1} & \cdots & a_1 x_{1,p} \\ \vdots & & \vdots \\ a_{n_1} x_{n_1,1} & \cdots & a_{n_1} x_{n_1,p} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \\ &= \frac{1}{\rho} \begin{pmatrix} r_1 \left(1 + \frac{r_1}{\lambda_1}\right) & \cdots & r_{n_1} \left(1 + \frac{r_{n_1}}{\lambda_{n_1}}\right) \end{pmatrix} \begin{pmatrix} a_1 \mathbf{x}_1^\top \boldsymbol{\beta} \\ \vdots \\ a_{n_1} \mathbf{x}_{n_1}^\top \boldsymbol{\beta} \end{pmatrix}. \end{aligned}$$

Next, consider the summation over Term 2,

$$\begin{aligned} \text{Term 2} &= \sum_{\{i:a_i=-1\}} \left(\frac{r_i}{\rho}\right)^2 \frac{1}{\lambda_i} (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2 \\ &= \left(\frac{1}{\rho^2}\right)\left(\frac{r_1^2}{\lambda_1}\right) a_1^2 (\mathbf{x}_1^\top \boldsymbol{\beta})^2 + \cdots + \left(\frac{1}{\rho^2}\right)\left(\frac{r_{n_1}^2}{\lambda_{n_1}}\right) a_{n_1}^2 (\mathbf{x}_{n_1}^\top \boldsymbol{\beta})^2. \end{aligned}$$

Observe that

$$
\begin{aligned}
&a_i^2\boldsymbol{\beta}^\top \mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\beta}\\
=&\boldsymbol{\beta}^\top \begin{pmatrix} a_i x_{i,1} \\ \vdots \\ a_i x_{i,p}\end{pmatrix}\begin{pmatrix} a_i x_{i,1} & \cdots & a_i x_{i,p}\end{pmatrix}\boldsymbol{\beta}\\
=&\boldsymbol{\beta}^\top \begin{pmatrix} a_i^2 x_{i,1}^2 & a_i^2 x_{i,1}x_{i,2} & \cdots & a_i^2 x_{i,1}x_{i,p} \\ a_i^2 x_{i,1}x_{i,2} & a_i^2 x_{i,2}^2 & \cdots & a_i^2 x_{i,2}x_{i,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_i^2 x_{i,1}x_{i,p} & a_i^2 x_{i,2}x_{i,p} & \cdots & a_i^2 x_{i,p}^2\end{pmatrix}\boldsymbol{\beta}\\
=&\begin{pmatrix}\beta_1 & \cdots & \beta_p\end{pmatrix}\begin{pmatrix} a_i^2 x_{i,1}^2 & a_i^2 x_{i,1}x_{i,2} & \cdots & a_i^2 x_{i,1}x_{i,p} \\ a_i^2 x_{i,1}x_{i,2} & a_i^2 x_{i,2}^2 & \cdots & a_i^2 x_{i,2}x_{i,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_i^2 x_{i,1}x_{i,p} & a_i^2 x_{i,2}x_{i,p} & \cdots & a_i^2 x_{i,p}^2\end{pmatrix}\begin{pmatrix}\beta_1 \\ \vdots \\ \beta_p\end{pmatrix}\\
=&\begin{pmatrix}\beta_1 a_i^2 x_{i,1}^2 + \cdots + \beta_p a_i^2 x_{i,1}x_{i,p} & \cdots & \beta_1 a_i^2 x_{i,1}x_{i,p} + \cdots + \beta_p a_i^2 x_{i,p}^2\end{pmatrix}\begin{pmatrix}\beta_1 \\ \vdots \\ \beta_p\end{pmatrix}\\
=&\beta_1^2 a_i^2 x_{i,1}^2 + \beta_1\beta_2 a_i^2 x_{i,1}x_{i,2} + \cdots + \beta_1\beta_p a_i^2 x_{i,1}x_{i,p}\\
&+ \cdots + \beta_1\beta_p a_i^2 x_{i,1}x_{i,p} + \beta_2\beta_p a_i^2 x_{i,2}x_{i,p} + \cdots + \beta_p^2 a_i^2 x_{i,p}^2\\
=&\sum_{j=1}^{p} a_i^2 x_{i,j}^2\beta_j^2 + 2\sum_{j=1}^{p}\sum_{k\neq j}^{p} a_i^2 x_{i,j}x_{i,k}\beta_j\beta_k\\
=&(a_i x_{i,1}\beta_1 + \cdots a_i x_{i,p}\beta_p)\cdot a_i(x_{i,1}\beta_1 + \cdots a_i x_{i,p}\beta_p)\\
=&(a_i\mathbf{x}_i^\top\boldsymbol{\beta})\cdot(a_i\mathbf{x}_i^\top\boldsymbol{\beta})\\
=&(a_i\mathbf{x}_i^\top\boldsymbol{\beta})^2.
\end{aligned}
$$

Also observe that

$$
\begin{aligned}
\boldsymbol{\beta}^\top\mathbf{X}_1^\top\mathbf{X}_1\boldsymbol{\beta} =&\boldsymbol{\beta}^\top \begin{pmatrix} a_1 x_{1,1} & \cdots & a_{n_1} x_{n_1,1} \\ \vdots & \ldots & \vdots \\ a_1 x_{1,p} & \cdots & a_{n_1} x_{n_1,p}\end{pmatrix}\begin{pmatrix} a_1 x_{1,1} & \cdots & a_1 x_{1,p} \\ \vdots & \ldots & \vdots \\ a_{n_1} x_{n_1,1} & \cdots & a_{n_1} x_{n_1,p}\end{pmatrix}\boldsymbol{\beta}\\
=&\begin{pmatrix} a_1\boldsymbol{\beta}^\top\mathbf{x}_1 & \cdots & a_{n_1}\boldsymbol{\beta}^\top\mathbf{x}_{n_1}\end{pmatrix}\begin{pmatrix} a_1\mathbf{x}_1^\top\boldsymbol{\beta} \\ \cdots \\ a_{n_1}\mathbf{x}_{n_1}^\top\boldsymbol{\beta}\end{pmatrix}\\
=&\begin{pmatrix} a_1^2\boldsymbol{\beta}^\top\mathbf{x}_1\mathbf{x}_1^\top\boldsymbol{\beta} \\ \vdots \\ a_{n_1}\boldsymbol{\beta}^\top\mathbf{x}_{n_1}\mathbf{x}_{n_1}^\top\boldsymbol{\beta}.\end{pmatrix}
\end{aligned}
$$

Define $\mathbf{\Lambda}_1 = diag(\lambda_1, \ldots, \lambda_{n_1})$. Then

$$\begin{aligned}
\text{Term 2} &= \left(\frac{1}{\rho^2}\right)\left(\frac{r_1^2}{\lambda_1}\right) a_1^2(\mathbf{x}_1^\top\boldsymbol{\beta})^2 + \cdots + \left(\frac{1}{\rho^2}\right)\left(\frac{r_{n_1}^2}{\lambda_{n_1}}\right) a_1^2(\mathbf{x}_{n_1}^\top\boldsymbol{\beta})^2 \\
&= \left(\frac{1}{\rho^2}\right)\begin{pmatrix}\left(\frac{r_1^2}{\lambda_1}\right) & \cdots & \left(\frac{r_{n_1}^2}{\lambda_{n_1}}\right)\end{pmatrix}\begin{pmatrix} a_1^2\boldsymbol{\beta}^\top\mathbf{x}_1\mathbf{x}_1^\top\boldsymbol{\beta} \\ \vdots \\ a_{n_1}^2\boldsymbol{\beta}^\top\mathbf{x}_{n_1}\mathbf{x}_{n_1}^\top\boldsymbol{\beta}\end{pmatrix} \\
&= \left(\frac{1}{\rho^2}\right)\begin{pmatrix}\left(\frac{r_1^2}{\lambda_1}\right) & \cdots & \left(\frac{r_{n_1}^2}{\lambda_{n_1}}\right)\end{pmatrix}\boldsymbol{\beta}^\top\mathbf{X}_1^\top\mathbf{X}_1\boldsymbol{\beta} \\
&= \boldsymbol{\beta}^\top\mathbf{X}_1^\top\mathbf{R}_1^\top\mathbf{\Lambda}_1^{-1}\mathbf{R}_1\mathbf{X}_1\boldsymbol{\beta}.
\end{aligned}$$

Let $n_{-1} = \sum_{i=1}^n \mathbb{1}(a_i = -1)$. Similarly, define $\mathbf{X}_{-1}$, $\mathbf{W}_{-1}$, $\mathbf{R}_{-1}$, and $\mathbf{\Lambda}_{-1}$ as

$$\mathbf{X}_{-1} \equiv \begin{pmatrix} a_1 x_{1,1} & \cdots & a_1 x_{1,p} \\ \vdots & & \vdots \\ a_{n_{-1}} x_{n_{-1},1} & \cdots & a_{n_{-1}} x_{n_{-1},p}\end{pmatrix}_{(n_{-1}\times p)}, \qquad \mathbf{W}_{-1} \equiv \begin{pmatrix} 1+\frac{r_1}{\lambda_1} \\ \vdots \\ 1+\frac{r_{n_{-1}}}{\lambda_{n_{-1}}}\end{pmatrix}_{(n_{-1}\times 1)},$$

$\mathbf{R}_{-1} \equiv diag(r_1/(1-\rho), \ldots, r_{n_{-1}}/(1-\rho))_{(n_{-1}\times n_{-1})}$, and $\mathbf{\Lambda}_{-1} = diag(\lambda_1, \ldots, \lambda_{n_{-1}})$.

Additionally define $\mathbf{\Sigma} \equiv diag(\sigma_1, \ldots, \sigma_p)$ so that we can write

$$\exp\left\{-\frac{1}{2}\sum_{j=1}^p \frac{\beta_j^2 - 2\beta_j\mu_{0,j}}{\sigma_0^2}\right\} = \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\mu}_0^\top\mathbf{\Sigma}^{-1}\boldsymbol{\beta})\right\}.$$

Thus, we have that

$$\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}) \\
&\propto \exp\left\{-\frac{1}{2}\left(-2\mathbf{W}_1^\top\mathbf{R}_1\mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}_1^\top\mathbf{R}_1^\top\mathbf{\Lambda}_1^{-1}\mathbf{R}_1\mathbf{X}_1\boldsymbol{\beta}\right)\right\} \\
&\quad \cdot \exp\left\{-\frac{1}{2}\left(-2\mathbf{W}_{-1}^\top\mathbf{R}_{-1}\mathbf{X}_{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}_{-1}^\top\mathbf{R}_{-1}^\top\mathbf{\Lambda}_{-1}^{-1}\mathbf{R}_{-1}\mathbf{X}_{-1}\boldsymbol{\beta}\right)\right\} \\
&\quad \cdot \exp\left\{-\frac{1}{2}(-2\boldsymbol{\mu}_0^\top\mathbf{\Sigma}^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\beta})\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^\top \underbrace{\left(\mathbf{X}_1^\top\mathbf{R}_1^\top\mathbf{\Lambda}_1^{-1}\mathbf{R}_1\mathbf{X}_1 + \mathbf{X}_{-1}^\top\mathbf{R}_{-1}^\top\mathbf{\Lambda}_{-1}^{-1}\mathbf{R}_{-1}\mathbf{X}_{-1} + \mathbf{\Sigma}^{-1}\right)}_{\equiv B_1^{-1}}\boldsymbol{\beta}\right.\right. \\
&\quad \left.\left. - 2(\underbrace{\mathbf{W}_1^\top\mathbf{R}_1\mathbf{X}_1 + \mathbf{W}_{-1}^\top\mathbf{R}_{-1}\mathbf{X}_{-1} + \boldsymbol{\mu}_0^\top\mathbf{\Sigma}^{-1}}_{\equiv b_1})\boldsymbol{\beta}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - B_1b_1)^\top B^{-1}(\boldsymbol{\beta} - B_1b_1) - b_1^\top B_1 b_1\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - B_1b_1)^\top B_1^{-1}(\boldsymbol{\beta} - B_1b_1)\right\}.
\end{aligned}$$

The conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$ is multivariate normal with mean B_1b_1 and variance-covariance matrix B_1.

5.6.3 Conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}$ (Exponential power prior)

$$
\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{r}, \mathbf{a}, \mathbf{x}) \\
&\quad \propto \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=1\}} \frac{\left(\frac{r_i}{\rho} + \lambda_i - \left(\frac{r_i}{\rho}\right) a_i \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\lambda_i} \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=-1\}} \frac{\left(\frac{r_i}{1-\rho} + \lambda_i - \left(\frac{r_i}{1-\rho}\right) a_i \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\lambda_i} \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2\nu^2} \sum_{j=1}^{p} \frac{\beta_j^2}{\sigma_j^2 \omega_j} \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=1\}} \left(-2\frac{r_i}{\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \left(1 + \frac{r_i}{\rho\lambda_i}\right) + \left(\frac{r_i}{\rho}\right)^2 \frac{1}{\lambda_i} (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right) \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2} \sum_{\{i:a_i=-1\}} \left(-2\frac{r_i}{1-\rho} a_i \mathbf{x}_i^\top \boldsymbol{\beta} \left(1 + \frac{1}{(1-\rho)\lambda_i}\right) + \left(\frac{r_i}{1-\rho}\right)^2 \frac{1}{\lambda_i} (a_i \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right) \right\} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2\nu^2} \sum_{j=1}^{p} \frac{\beta_j^2}{\sigma_j^2 \omega_j} \right\}
\end{aligned}
$$

The summation inside the first and second exponential terms are the same as in the derivation under the normal distribution prior for $\boldsymbol{\beta}$ (Section 5.6.2). Letting $\boldsymbol{\Omega} \equiv diag(\omega_1, \ldots, \omega_p)_{(p\times p)}$, we can write

$$
\begin{aligned}
\sum_{j=1}^{p} \frac{\beta_j^2}{\sigma_j^2 \omega_j} &= \begin{pmatrix} 1/(\sigma_1^2\omega_1) & \cdots & 1/(\sigma_p^2\omega_p) \end{pmatrix} \begin{pmatrix} \beta_1^2 \\ \vdots \\ \beta_p^2 \end{pmatrix} \\
&= \begin{pmatrix} \beta_1^2 & \cdots & \beta_p^2 \end{pmatrix} \begin{pmatrix} 1/(\sigma_1^2\omega_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/(\sigma_p^2\omega_p) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \\
&= \begin{pmatrix} \beta_1^2 & \cdots & \beta_p^2 \end{pmatrix} \begin{pmatrix} 1/\sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_p^2 \end{pmatrix} \begin{pmatrix} 1/\omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\omega_p \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \\
&= \boldsymbol{\beta}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}
\end{aligned}
$$

Thus, we have that

$$
\begin{aligned}
&p(\boldsymbol{\beta}|\boldsymbol{\lambda},\boldsymbol{\omega},\mathbf{r},\mathbf{a},\mathbf{x}) \\
&\quad\propto \exp\left\{-\frac{1}{2}\left(-2\mathbf{W}_1^\top\mathbf{R}_1\mathbf{X}_1\boldsymbol{\beta}+\boldsymbol{\beta}^\top\mathbf{X}_1^\top\mathbf{R}_1^\top\boldsymbol{\Lambda}_1^{-1}\mathbf{R}_1\mathbf{X}_1\boldsymbol{\beta}\right)\right\} \\
&\qquad\cdot \exp\left\{-\frac{1}{2}\left(-2\mathbf{W}_{-1}^\top\mathbf{R}_{-1}\mathbf{X}_{-1}\boldsymbol{\beta}+\boldsymbol{\beta}^\top\mathbf{X}_{-1}^\top\mathbf{R}_{-1}^\top\boldsymbol{\Lambda}_{-1}^{-1}\mathbf{R}_{-1}\mathbf{X}_{-1}\boldsymbol{\beta}\right)\right\} \\
&\qquad\cdot \exp\left\{-\frac{1}{2}\nu^{-2}\boldsymbol{\beta}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right\} \\
&\quad= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^\top\underbrace{\left(\mathbf{X}_1^\top\mathbf{R}_1^\top\boldsymbol{\Lambda}_1^{-1}\mathbf{R}_1\mathbf{X}_1+\mathbf{X}_{-1}^\top\mathbf{R}_{-1}^\top\boldsymbol{\Lambda}_{-1}^{-1}\mathbf{R}_{-1}\mathbf{X}_{-1}+\nu^{-2}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}^{-1}\right)}_{\equiv B_2^{-1}}\boldsymbol{\beta}\right.\right. \\
&\qquad\left.\left.-2(\underbrace{\mathbf{W}_1^\top\mathbf{R}_1\mathbf{X}_1+\mathbf{W}_{-1}^\top\mathbf{R}_{-1}\mathbf{X}_{-1}}_{\equiv b_2})\boldsymbol{\beta}\right]\right\} \\
&\quad= \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-B_2b_2)^\top B_2^{-1}(\boldsymbol{\beta}-B_2b_2)-b_2^\top B_2b_2\right\} \\
&\quad\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-B_2b_2)^\top B_2^{-1}(\boldsymbol{\beta}-B_2b_2)\right\}.
\end{aligned}
$$

The conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$, $\boldsymbol{\omega}$, and is multivariate normal with mean B_2b_2 and variance-covariance matrix B_2.

Chapter 6

Conclusion and perspectives

We explored the concept of decision rules in precision medicine by focusing on two methods for constructing optimal rules: Q-learning and Outcome-Weighted Learning (OWL), with a particular emphasis on integrating expert knowledge. The first method, Q-learning, led us to examine decision-making processes within the framework of Dynamic Treatment Regimes (DTR) and their broader connection to Reinforcement Learning (RL), while considering the specificities of observational data. This analysis served as the foundation for a comprehensive review on integrating medical expertise within this context, which subsequently guided the development of a method for incorporating expert knowledge into RL models through preference-based reward learning. The second method, OWL applied to Individualized Treatment Regime (ITR), led us to reconsider its formalism from a Bayesian perspective, allowing us to quantify uncertainty and provide practitioners with a new decision support tool.

Our exploration begins with the specific characteristics of the data, which impose certain constraints on the application of RL. First, the modeling of RL traditionally relies on the Markov assumption, which presumes that the current state contains all the necessary information to make an optimal decision. However, this assumption often proves too restrictive in practice. To overcome this limitation, it is possible to consider the patient's entire history, but this approach demands significant computational resources and extensive data, which in practice limits the application of DTRs to a reduced number of decision points. Second, when applying RL to medical data, the issue of causality becomes crucial, particularly with observational data, which is commonly used for DTRs. To better understand the challenges of causality in precision medicine, we first examined this issue within the simpler context of ITR, which allowed us to illustrate the key assumptions governing this problem. These assumptions can be generalized to the DTR framework. However, it is important to note that RL does not inherently address causality, even though there has been considerable research on this topic [11, 140]. Addressing the causality issue, while using RL, largely depends on the data and how it is collected. A clinical trial design that addresses this issue is the sequential multiple assignment randomized trials, which are specifically designed to satisfy all causal assumptions. However, such clinical trials are expensive and difficult to implement, leading to a limited number of available datasets. Third, the choice of RL algorithms for DTRs must meet specific criteria: they must be model-free, value-based, and off-policy to ensure their effectiveness in an offline context. Among these algorithms, backward Q-learning, or Fitted Q-Iteration, stands out as the most commonly used in RL applications for DTRs.

The first contribution of this thesis is the development of a state-of-the-art review

on the integration of medical knowledge into RL models. A key observation, based on the fact that data generally comes from observational studies, reveals that this restricts the application of methods to offline contexts, with specific algorithmic properties needed for such settings. We then identified the properties of each algorithm and compared them with the desired characteristics for effective use in the context of DTRs. Beyond these technical considerations, the application of RL in clinical settings may face resistance due to the complexity and interpretability of the proposed treatment strategies. A key issue is the acceptability of the optimal DTR by both patients and practitioners, which largely depends on the understanding of the decision rules. Therefore, integrating medical expertise into machine learning methods for personalized treatments is essential to enhance safety, interpretability, and effectiveness. The integration of this expertise can occur at various stages of the RL process or within its key components, such as rewards, value functions, the objective function, or the policy. First, medical knowledge is often integrated at the outset of the study, during the design of the experiment. Physicians contribute to the selection of variables used for learning the decision rule. Similarly, the selection of algorithms is made in collaboration between medical experts and machine learning experts, depending on the application framework and available data. Second, medical knowledge can be incorporated by influencing the rewards, which are one of the key elements of an RL algorithm since they directly impact the construction of the decision rule. Therefore, their design is crucial. Traditionally, a variable representative of the study's objective is chosen to define the rewards. However, some methods propose a less ad hoc approach to designing these rewards. Among these methods, inverse reinforcement learning [104, 68] and preference learning [27, 2] seek to leverage expert knowledge from a dataset or preferences to generalize the construction of rewards. Human-centered RL [56] directly replaces rewards with expert feedback. However, preference learning and human-centered RL are often limited to interactive contexts, making them inapplicable to DTRs and observational clinical applications. In contrast, inverse reinforcement learning is promising as it is developed in an offline context and is well-suited for real clinical applications. Third, the learning of decision rules can be achieved through value functions, allowing the integration of medical expertise at this level. One approach is to incorporate observed medical mechanisms; specifically, it involves penalizing Q-values associated with non-decisive treatments [29]. However, this method was initially developed in an online context and needs to be reassessed for offline settings. Another idea is to establish a relay between human decisions and those proposed by the algorithm. In one scenario, the physician would take over when the patient is in a critical situation [128]. In another, the algorithm would suggest alternative treatments to those traditionally proposed, along with associated risks [109]. These hybrid methods seem promising for real clinical applications, but concrete evidence of their implementation is still lacking. In the policy-based methodological framework, expertise integration can be achieved through a method called supervised RL [134, 124]. Its goal is to faithfully replicate common medical practices, offering a precise emulation of physicians' decisions. However, it does not allow for the discovery of alternative or underexplored treatments compared to conventional care methods. Finally, decision rule learning can

be approached methodologically through the policy. It is important to note that classical RL methods typically recommend only one policy, meaning one treatment and one dose for each decision time. To enrich the context, multi-policies methods have been developed to offer an expert multiple equivalent treatments to choose from. The work of [65] is particularly suitable for application to DTRs based on observational data, but it was developed within a framework of patient preferences and could be reassessed within an expert preference framework.

This first part of our resarches on both RL for DTR and the state-of-the-art review on the integration of expert knowledge is covered in an article available on arXiv:2407.00364, which will be submitted for publication soon. In my future research, I plan to delve deeper into methods related to multiple policies, with the aim of offering practitioners not just a single optimal treatment, but multiple equivalent treatment options. The concept of near-optimality is particularly complex in a multistage context, as choosing an action that is close in equivalence may lead to a gradual deviation from optimality over time.

The second contribution of this thesis directly stems from the previous state-of-the-art review and the conclusions drawn from it. We were particularly interested in the generalization of reward construction, especially through preference learning. We thus developed a reward learning method based on preference learning, specifically designed for application to DTRs. This process unfolds in three steps: (1) an expert expresses preferences between pairs of elements, which induces a ranking among all instances in the previously collected dataset; (2) rewards are then constructed using a Bradley-Terry probabilistic model; (3) these rewards are used to learn the policy in Q-learning models. The main strength of this method lies in its ability to construct rewards in a generalized, data-driven manner. It leverages both the expertise of healthcare professionals and the relationships between patient data, thereby avoiding manual reward construction that can be arbitrary, while ensuring consistency in the learning of medical strategies. This method was illustrated by two case studies: one on the treatment of adolescents with obesity [8, 61], and the other on a cancer simulation [144]. In our first case study, the objective was to demonstrate that rewards constructed using a preference model, whether based on stages or trajectories, capture the same variations and dispersions as the rewards observed in the classical approach. However, our method stands out particularly in the second case study, where we examined the strategies learned from different models. The presented method, which generates rewards from a data-driven preference model, offers a more generalized approach compared to the initial model, where rewards were defined more subjectively. Although the distributions of rewards generated by our model and those of the initial model are different and weakly correlated, the learned strategies produce the expected medical results, effectively balancing treatment toxicity and tumor size. However, our work has some limitations and could benefit from improvements. The estimation of rewards from the Bradley-Terry pairwise comparison model relies on the Newton-Raphson algorithm. The literature suggests that the minorization-maximization algorithm [42] could be a more efficient estimation technique. Additionally, other comparison models, stemming from research in social choice or sports statistics, could also be explored, such as the Thurstone-Mosteller

model [34], the Elo model [14], and the Plackett-Luce model [87]. Another limitation of our method is that it directly provides reward values from comparisons. It would be interesting to reformulate these models into a parametric reward function based on state variables.

This second contribution will be the subject of a forthcoming article. Moving forward, I would like to explore the last limitation identified: the parametric formulation of the model into a function dependent on patient states. This would enable the development of an interpretable and explainable reward function, highlighting the importance of variables in the construction of decision rules.

The third contribution of this thesis is the development of a Bayesian outcome-weighted learning method aimed at quantifying uncertainty in treatment recommendations. By introducing a Bayesian framework, this approach enhances traditional OWL methods by enabling a probabilistic estimation of decision rules. By reformulating the optimization problem within a probabilistic framework, our method generates a complete posterior distribution, offering both inferential capabilities and a precise evaluation of the uncertainty associated with the recommended treatments. The main steps of our contribution are as follows. First, we propose a Bayesian approach for learning optimal ITR, based on a classification framework. Next, we developed a simple Gibbs sampling algorithm to facilitate this learning process. Finally, we demonstrate how the resulting pseudo-posterior distribution can be used to quantify uncertainty in treatment recommendations, with performance illustrated through simulation studies. This ability to assess uncertainty on an individual basis is particularly useful for the design and implementation of clinical studies, by identifying patients for whom recommendations are reliable and those requiring additional information. However, our work has some limitations. We focused solely on linear rules, which, while simple to interpret, can sometimes be outperformed by nonlinear rules that offer significant clinical improvements, even if they are less interpretable. An extension to nonlinear rules could draw on work related to Bayesian support vector machines with nonlinear decision boundaries, such as those proposed by [36]. Additionally, our method does not incorporate variable selection, which limits its application in high-dimensional contexts. Strategies such as L1 regularization or the use of spike-and-slab priors, as suggested by [88], could be explored to overcome this limitation.

This third contribution is the result of joint work with Nikki L. B. Freeman, accessible on arXiv:2406.115. Before submission, we plan to further enhance the work by including an application on observational data. Moving forward, we particularly aim to extend Bayesian OWL to a nonlinear framework, which would more accurately reflect decision rules on observational data, and incorporate variable selection to identify and retain only the most relevant variables for the model, eliminating those that are redundant or uninformative, as highlighted in the model's limitations.

The research presented, which includes both RL methods and OWL approach, focuses on enhancing the personalization of care by integrating clinical specificities and observational data. These strategies aim to tailor treatments more effectively to individual patient characteristics while incorporating medical expertise at various

stages of the process. By embedding medical knowledge early on, it becomes possible to construct rewards that fully utilize data and expert insights. Additionally, the Bayesian OWL method introduces a new tool for analyzing optimal strategies through uncertainty quantification. Overall, this contributions offer promising directions for making machine learning algorithms more relevant and better suited to the challenges of precision medicine, potentially increasing practitioners' trust in these tools.

This work also highlights the importance of interdisciplinary collaboration, particularly among the fields of machine learning, medical sciences, and statistics. The development of the proposed methods, which integrate medical knowledge into reinforcement learning models and quantify uncertainty in treatment recommendations, could greatly benefit from close interaction among researchers from these areas. Such collaborations are essential for ensuring that the algorithms developed meet specific clinical needs and remain comprehensible to healthcare professionals. Moreover, they are necessary to guarantee that the proposed methods are well-suited to clinical contexts and genuinely practical. Integrating expertise from various disciplines could also facilitate the adoption of these technologies in clinical settings, offering solutions that are both technically robust and clinically relevant.

Appendix

Table : Reinforcement learning applications for sequential decision in healthcare

Table 6.1 – **Reinforcement learning applications for sequential decision in healthcare.** Ref, References; Environment, description of the medical application context; Data, Simulated or Real; Model, Decision Process (DP) or Markov Decision Process (MPD) or Partially Observable Markov Decision Process (POMDP); Stage, number of stages; Off/On, offline or online; Algorithm, standard reference algorithm of reinforcement learning without the paper specifications or innovation added.

Ref.	Environment	Data	Model	Stage	Off/On	Algorithm
[29]	Simulated patient with anemia due to kidney failure	Real Data	MDP	24	Online	Q-learning
[144]	ODE Simulation of cancer trial for advanced generic cancer of treatment	Simulation	DP	6	Offline	Backward Q-learning
[35]	ODE Simulation of cancer growth on a cell population level	Simulation	MDP	N/A	Online	Q-learning
[1]	ODE Simulation of cancer growth on a cell population level	Simulation	DP	N/A	Online	Actor-Critic
[107]	CATIE	Real Data	DP	2	Offline	Backward Q-learning
[2]	Same as in [144]	Simulation	MDP	12	Online	Policy Search
[32]	Same as in [144]	Simulation	DP	3	Offline	Backward Q-learning
[79]	ADHD	Real Data	DP	2	Offline	Backward Q-learning
[27]	Same as in [144]	Simulation	MDP	6	Online	Policy Iteration
[145]	Exponential distribution simulation of cancer for parties in phase III.	Simulation	DP	2	Offline	Backward Q-learning
[123]	(Chapter 4) Electrical stimulation for epilepsy from in vitro experiments	Real Data	MDP	N/A	Offline	Backward Q-learning
[123]	(Chapter 5) Electrical stimulation for Parkinson's disease	Simulation	MDP	2, 6, 9, 10	Online	SARSA, Temporal Difference
[123]	(Chapter 5) Electrical stimulation for Parkinson's disease	Simulation	MDP	2, 6, 9, 10	Offline	Backward Q-learning
[123]	(Chapter 6) Fractionation scheduling for radiation therapy	Simulation	MDP	4, 10, 30	Offline	Backward Q-learning
[54]	ADHD	Real Data	DP	2	Offline	Backward Q-learning
[53]	STAR*D	Real Data	DP	2	Offline	Backward Q-learning
[23]	Simulated patients with diabetes	Simulation	MDP	Indefinite	Online	Q-learning

Continued on next page

Table 6.1 – Continued from previous page

Ref.	Environment	Data	Model	Stage	Off/On	Algorithm
[13]	Comparison of depression interventions after acute coronary syndrom (SMART)	Real Data	DP	2	Offline	Backward Q-learning
[64]	STAR*D and ADHD	Real Data	DP	2	Offline	Outcome-Weighted Learning and Q-learning
[41]	Same as in [144]	Simulation	MDP	N/A	Offline	Backward Q-learning
[85]	ODE Simulation of cancer growth on a cell population level	Simulation	MDP	N/A	Online	Q-learning
[118]	114 patients used to construct synthetic data by GAN	Simulation	MDP	35	Online	Deep Q-learning
[44]	Model of tumor growth using NetLogo package, Agent-based simulation	Simulation	DP	N/A	Online	Q-learning
[62]	Registry data from 6021 AML patients who underwent allogeneic stem cell transplantation	Real Data	DP	5	Offline	Deep Q-learning
[52]	Nonrandomized registry data from 11,141 patients who underwent allogeneic stem cell transplantation	Real Data	DP	2	Offline	Backward Q-learning
[90]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning
[89]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning
[130]	Linear and ODE Simulation of cancer trial	Simulation	MDP	N/A	Online	Deep Q-learning
[86]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning
[131]	ODE Simulation of cancer growth on a cell population level	Simulation	MDP	N/A	Online	Q-learning
[67]	Simulated patients with diabetes	Simulation	MDP	Indefinite	Online	V-learning
[58]	MIMIC	Real Data	POMDP	N/A	Offline	Deep Q-Learning
[137]	ODE Simulation of cancer growth on a cell population level	Simulation	MDP	N/A	Online	Q-learning
[17]	Real dataset of breast cancer	Real Data	MDP	N/A	Online	Q-learning
[114]	MIMIC	Real Data	MDP	750	Offline	Temporal Difference
[109]	MIMIC	Real Data	MDP	N/A	Offline	Q-learning
[74]	Simulate tumor development inside healthy tissue and the effect of radiation therapy	Simulation	MDP	N/A	Online	Deep Policy Gradient
[128]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning
[106]	40 patients of stage-four colon cancer	Real Data	MDP	6	Offline	Deep Q-Learning
[48]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning

Bibliography

[1] Inkyung Ahn and Jooyoung Park. "Drug Scheduling of Cancer Chemotherapy Based on Natural Actor-Critic Approach". In: *BioSystems* 106.2-3 (2011), pp. 121–129.

[2] Riad Akrour, Marc Schoenauer, and Michèle Sebag. "APRIL: Active Preference Learning-Based Reinforcement Learning". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012*. 2012, pp. 116–131.

[3] Yorghos Apostolopoulos, Kristen Hassmiller Lich, and Michael K Lemke. *Complex Systems and Population Health*. Oxford University Press, 2020.

[4] Elja Arjas and Olli Saarela. "Optimal Dynamic Regimes: Presenting a Case for Predictive Inference". In: *International Journal of Biostatistics* 6.2 (2010), Article 10.

[5] Christian Arzate Cruz and Takeo Igarashi. "A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges". In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 2020, pp. 1195–1209.

[6] Richard Bellman. "A Markovian Decision Process". In: *Journal of Mathematics and Mechanics* (1957), pp. 679–684.

[7] Richard Bellman. "Dynamic Programming". In: *Science* 153.3731 (1966), pp. 34–37.

[8] Robert I Berkowitz, Thomas A Wadden, Christine A Gehrman, Chanelle T Bishop-Gilyard, Reneé H Moore, Leslie G Womble, Joanna L Cronquist, Natalie L Trumpikas, Lorraine E Levitt-Katz, and Melissa S Xanthopoulos. "Meal Replacements in the Treatment of Adolescent Obesity: a Randomized Controlled Trial". In: *Obesity* 19.6 (2011), pp. 1193–1199.

[9] Dimitri P Bertsekas. *Dynamic programming: deterministic and stochastic models*. Prentice-Hall, Inc., 1987.

[10] Ralph Allan Bradley and Milton E Terry. "Rank Analysis of Incomplete Block Designs: I. The method of Paired Comparisons". In: *Biometrika* 39.3/4 (1952), pp. 324–345.

[11] Bibhas Chakraborty and Susan A Murphy. "Dynamic Treatment Regimes". In: *Annual Review of Statistics and its Application* 1 (2014), pp. 447–464.

[12] Nicholas C Chesnaye, Vianda S Stel, Giovanni Tripepi, Friedo W Dekker, Edouard L Fu, Carmine Zoccali, and Kitty J Jager. "An Introduction to Inverse Probability of Treatment Weighting in Observational Research". In: *Clinical Kidney Journal* 15.1 (2022), pp. 14–20.

[13] Ying Kuen Cheung, Bibhas Chakraborty, and Karina W Davidson. "Sequential Multiple Assignment Randomized Trial (SMART) with Adaptive Randomization for Quality Improvement in Depression Treatment Program". In: *Biometrics* 71.2 (2015), pp. 450–459.

[14] Andrew P Clark, Kate L Howard, Andy T Woods, Ian S Penton-Voak, and Christof Neumann. "Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength". In: *PloS one* 13.1 (2018).

[15] Jesse Clifton and Eric Laber. "Q-learning: Theory and Applications". In: *Annual Review of Statistics and Its Application* 7 (2020), pp. 279–301.

[16] Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. "Reinforcement learning for Intelligent Healthcare Applications: A Survey". In: *Artificial Intelligence in Medicine* 109 (2020).

[17] Salma Daoud, Afef Mdhaffar, Mohamed Jmaiel, and Bernd Freisleben. "Q-rank: Reinforcement Learning for Recommending Algorithms to Predict Drug Sensitivity to Cancer Therapy". In: *IEEE Journal of Biomedical and Health Informatics* 24.11 (2020), pp. 3154–3161.

[18] Nina Deliu and Bibhas Chakraborty. "Dynamic Treatment Regimes for Optimizing Healthcare". In: *The Elements of Joint Learning and Optimization in Operations Management*. Springer, 2022, pp. 391–444.

[19] Nina Deliu, Joseph Jay Williams, and Bibhas Chakraborty. "Reinforcement Learning in Modern Biostatistics: Constructing Optimal Adaptive Interventions". In: *International Statistical Review* (2024).

[20] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer New York, 1986.

[21] Jan-Niklas Eckardt, Karsten Wendt, Martin Bornhaeuser, and Jan Moritz Middeke. "Reinforcement Learning for Precision Oncology". In: *Cancers* 13.18 (2021), p. 4624.

[22] Damien Ernst, Pierre Geurts, and Louis Wehenkel. "Tree-Based Batch Mode Reinforcement Learning". In: *Journal of Machine Learning Research* 6 (2005), pp. 503–556.

[23] Ashkan Ertefaie and Robert L Strawderman. "Constructing Dynamic Treatment Regimes over Indefinite Time Horizons". In: *Biometrika* 105.4 (2018), pp. 963–977.

[24] Yuxin Fan. "Value Search Estimators of Individualized Treatment Regimes Using a New Class of Weights". PhD thesis. 2016.

[25] David Firth and Heather Turner. "Bradley-Terry models in R: the BradleyTerry2 package". In: *Journal of Statistical Software* 48.9 (2012).

[26] Sheng Fu, Qinying He, Sanguo Zhang, and Yufeng Liu. "Robust Outcome Weighted Learning for Optimal Individualized Treatment Rules". In: *Journal of Biopharmaceutical Statistics* 29.4 (2019), pp. 606–624.

[27] Johannes Fürnkranz, Eyke Hüllermeier, Weiwei Cheng, and Sang-Hyeun Park. "Preference-Based Reinforcement Learning: a Formal Framework and a Policy Iteration Algorithm". In: *Machine Learning* 89 (2012), pp. 123–156.

[28] Frédérick Garcia and Emmanuel Rachelson. "Markov Decision Processes". In: *Markov Decision Processes in Artificial Intelligence* (2013), pp. 1–38.

[29] Adam E Gaweda, Mehmet K Muezzinoglu, George R Aronoff, Alfred A Jacobs, Jacek M Zurada, and Michael E Brier. "Incorporating Prior Knowledge into Q-Learning for Drug Delivery Individualization". In: *Fourth International Conference on Machine Learning and Applications (ICMLA'05)*. IEEE. 2005, 6–pp.

[30] Edward I George and Robert E McCulloch. "Variable Selection via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889.

[31] C Taylor Gilliland, Julia White, Barry Gee, Rosan Kreeftmeijer-Vegter, Florence Bietrix, Anton E Ussi, Marian Hajduch, Petr Kocis, Nobuyoshi Chiba, Ryutaro Hirasawa, et al. *The Fundamental Characteristics of a Translational Scientist*. 2019.

[32] Yair Goldberg and Michael R Kosorok. "Q-learning with Censored Data". In: *Annals of statistics* 40.1 (2012), p. 529.

[33] Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert Babuska. "A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients". In: *IEEE Transactions on Systems, Man, and Cybernetics, part C (applications and reviews)* 42.6 (2012), pp. 1291–1307.

[34] John C Handley. "Comparative Analysis of Bradley-Terry and Thurstone-Mosteller Paired Comparison Models for Image Quality Assessment". In: *PICS*. Vol. 1. 2001, pp. 108–112.

[35] Amin Hassani et al. "Reinforcement Learning Based Control of Tumor Growth with Chemotherapy". In: *2010 International Conference on System Science and Engineering*. IEEE. 2010, pp. 185–189.

[36] Ricardo Henao, Xin Yuan, and L Carin. "Bayesian Nonlinear Support Vector Machines and Discriminative Factor Modeling". In: *Advances in Neural Information Processing Systems* (2014), pp. 1754–1762.

[37] M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2023.

[38] Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Vol. 30. Springer Science & Business Media, 2012.

[39] Andreas Holzinger. "Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop?" In: *Brain Informatics* 3.2 (2016), pp. 119–131.

[40] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. "Causability and Explainability of Artificial Intelligence in Medicine". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4 (2019), e1312.

[41] Kyle Humphrey. "Using reinforcement learning to personalize dosing strategies in a simulated cancer trial with high dimensional data". PhD thesis. 2017.

[42] David R Hunter. "MM Algorithms for Generalized Bradley-Terry Models". In: *The Annals of Statistics* 32.1 (2004), pp. 384–406.

[43] Robert Istepanian, Swamy Laxminarayan, and Constantinos S Pattichis. *M-health: Emerging Mobile Health Systems*. Springer Science & Business Media, 2007.

[44] Ammar Jalalimanesh, Hamidreza Shahabi Haghighi, Abbas Ahmadi, and Madjid Soltani. "Simulation-Based Optimization of Radiotherapy: Agent-Based Modeling and Reinforcement Learning". In: *Mathematics and Computers in Simulation* 133 (2017), pp. 235–248.

[45] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. "MIMIC-III: A Freely Accessible Critical Care Database". In: *Scientific Data* 3.1 (2016), pp. 1–9.

[46] Anna R Kahkoska, Nikki LB Freeman, and Kristen Hassmiller Lich. "Systems-Aligned Precision Medicine—Building an Evidence Base for Individuals Within Complex Systems". In: *JAMA Health Forum*. Vol. 3. 7. American Medical Association. 2022.

[47] Anna R Kahkoska, Kristen Hassmiller Lich, and Michael R Kosorok. "Focusing on Optimality for the Translation of Precision Medicine". In: *Journal of Clinical and Translational Science* 6.1 (2022).

[48] Pramod Kaushik, Sneha Kummetha, Perusha Moodley, and Raju S. Bapi. *A Conservative Q-Learning Approach for Handling Distribution Shift in Sepsis Treatment Strategies*. 2022. arXiv: 2203.13884 [cs.LG].

[49] Matthieu Komorowski, A Gordon, LA Celi, and A Faisal. "A Markov Decision Process to Suggest Optimal Treatment of Severe Infections in Intensive Care". In: *Neural Information Processing Systems Workshop on Machine Learning for Health*. 2016.

[50] Michael R Kosorok and Eric B Laber. "Precision Medicine". In: *Annual Review of Statistics and its Application* 6 (2019), pp. 263–286.

[51] Michael R Kosorok and Erica EM Moodie. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. SIAM, 2015.

[52] Elizabeth F Krakow, Michael Hemmer, Tao Wang, Brent Logan, Mukta Arora, Stephen Spellman, Daniel Couriel, Amin Alousi, Joseph Pidala, Michael Last, et al. "Tools for the Precision Medicine Era: How to Develop Highly Personalized Treatment Recommendations from Cohort and Registry Data Using Q-Learning". In: *American journal of epidemiology* 186.2 (2017), pp. 160–172.

[53] Eric B Laber, Kristin A Linn, and Leonard A Stefanski. "Interactive Model Building for Q-Learning". In: *Biometrika* 101.4 (2014), pp. 831–847.

[54] Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, and Susan A Murphy. "Dynamic Treatment Regimes: Technical Challenges and Applications". In: *Electronic Journal of Statistics* 8.1 (2014), p. 1225.

[55] Eric B. Laber and Ana-Maria Staicu. "Functional Feature Construction for Individualized Treatment Regimes". In: *Journal of the American Statistical Association* 113.523 (2017), pp. 1219–1227.

[56] Guangliang Li, Randy Gomez, Keisuke Nakamura, and Bo He. "Human-Centered Reinforcement Learning: A Survey". In: *IEEE Transactions on Human-Machine Systems* 49.4 (2019), pp. 337–349.

[57] Luchen Li, Ignacio Albert-Smet, and Aldo A. Faisal. *Optimizing Medical Treatment for Sepsis in Intensive Care: from Reinforcement Learning to Pre-Trial Evaluation*. 2020. arXiv: 2003.06474 [cs.LG].

[58] Luchen Li, Matthieu Komorowski, and Aldo A. Faisal. *Optimizing Sequential Medical Treatments with Auto-Encoding Heuristic Search in POMDPs*. 2019. arXiv: 1905.07465 [cs.AI].

[59] Luchen Li, Matthieu Komorowski, and Aldo A. Faisal. *The Actor Search Tree Critic (ASTC) for Off-Policy POMDP Learning in Medical Decision Making*. 2018. arXiv: 1805.11548 [cs.AI].

[60] Zhen Li, Jie Chen, Eric Laber, Fang Liu, and Richard Baumgartner. "Optimal Treatment Regimes: A Review and Empirical Comparison". In: *International Statistical Review* 91.3 (2023), pp. 427–463.

[61] Kristin A Linn, Eric B Laber, and Leonard A Stefanski. "iqLearn: Interactive Q-learning in R". In: *Journal of Statistical Software* 64.1 (2015).

[62] Ying Liu, Brent Logan, Ning Liu, Zhiyuan Xu, Jian Tang, and Yangzhi Wang. "Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data". In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE. 2017, pp. 380–385.

[63] Ying Liu, Yuanjia Wang, Michael R Kosorok, Yingqi Zhao, and Donglin Zeng. "Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens". In: *Statistics in Medicine* 37.26 (2018), pp. 3776–3788.

[64] Ying Liu, Yuanjia Wang, Michael R. Kosorok, Yingqi Zhao, and Donglin Zeng. *Robust Hybrid Learning for Estimating Personalized Dynamic Treatment Regimens*. 2016. arXiv: 1611.02314 [stat.ME].

[65] Daniel J Lizotte and Eric B Laber. "Multi-Objective Markov Decision Processes for Data-Driven Decision Support". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 7378–7405.

[66] Tamlin Love, Ritesh Ajoodha, and Benjamin Rosman. "Who Should I Trust? Cautiously Learning with Unreliable Experts". In: *Neural Computing and Applications* 35.23 (2023), pp. 16865–16875.

[67] Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. "Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning". In: *Journal of the American Statistical Association* (2019).

[68] Daniel J Luckett, Eric B Laber, Siyeon Kim, and Michael R Kosorok. "Estimation and Optimization of Composite Outcomes". In: *Journal of Machine Learning Research* 22.167 (2021), pp. 1–40.

[69] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. "A Review on Human-AI Interaction in Machine Learning and Insights for Medical Applicationss". In: *International Journal of Environmental Research and Public Health* 18.4 (2021), p. 2121.

[70] M. Milani Fard and J. Pineau. "Non-Deterministic Policies in Markovian Decision Processes". In: *Journal of Artificial Intelligence Research* 40 (2011), pp. 1–24.

[71] T J Mitchell and J J Beauchamp. "Bayesian Variable Selection in Linear Regression". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

[72] George E Monahan. "State of the Art—A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms". In: *Management Science* 28.1 (1982), pp. 1–16.

[73] Erica E M Moodie, Bibhas Chakraborty, and Michael S Kramer. "Q-Learning for Estimating Optimal Dynamic Treatment Rules from Observational Data". In: *Canadian Journal of Statistics* 40.4 (2012), pp. 629–645.

[74] Grégoire Moreau, Vincent François-Lavet, Paul Desbordes, and Benoıt Macq. "Reinforcement Learning for Radiotherapy Dose Fractioning Automation". In: *Biomedicines* 9.2 (2021), p. 214.

[75] Susan A Murphy. "A Generalization Error for Q-Learning". In: *Journal of Machine Learning Research* 6 (2005), pp. 1073–1097.

[76] Susan A Murphy. "An Experimental Design for the Development of Adaptive Treatment Strategies". In: *Statistics in Medicine* 24.10 (2005), pp. 1455–1481.

[77] Susan A Murphy. "Optimal Dynamic Treatment Regimes". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 65.2 (2003), pp. 331–355.

[78] Thomas A Murray, Ying Yuan, and Peter F Thall. "A Bayesian Machine Learning Approach for Optimizing Dynamic Treatment Regimes". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1255–1267.

[79] Inbal Nahum-Shani, Min Qian, Daniel Almirall, William E Pelham, Beth Gnagy, Gregory A Fabiano, James G Waxmonsky, Jihnhee Yu, and Susan A Murphy. "Q-Learning: A Data Analysis Method for Constructing Adaptive Interventions". In: *Psychological Methods* 17.4 (2012), p. 478.

[80] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. "Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support". In: *Annals of Behavioral Medicine* (2018), pp. 1–17.

[81] Leland Gerson Neuberg. "Causality: Models, Reasoning, and Inference, by Judea Pearl, Cambridge University Press, 2000". In: *Econometric Theory* 19.4 (2003), pp. 675–685.

[82] Christophe Nivot. "Analyse et Étude des Processus Markoviens Décisionnels". PhD thesis. Bordeaux, 2016.

[83] Liliana Orellana, Andrea Rotnitzky, and James M Robins. "Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content". In: *The International Journal of Biostatistics* 6.2 (2010), Article 8.

[84] Dirk Ormoneit and Śaunak Sen. "Kernel-Based Reinforcement Learning". In: *Machine Learning* 49 (2002), pp. 161–178.

[85] Regina Padmanabhan, Nader Meskin, and Wassim M Haddad. "Reinforcement Learning-Based Control of Drug Dosing for Cancer Chemotherapy Treatment". In: *Mathematical Biosciences* 293 (2017), pp. 11–20.

[86] Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, H Lehman Li-wei, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. "Improving Sepsis Treatment Strategies by Combining Deep and Kernel-Based Reinforcement Learning". In: *AMIA Annual Symposium Proceedings*. Vol. 2018. American Medical Informatics Association. 2018, p. 887.

[87] R. L. Plackett. "The Analysis of Permutations". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24.2 (1975), pp. 193–202.

[88] Nicholas G Polson and Steven L Scott. "Data Augmentation for Support Vector Machines". In: *Bayesian Analysis* 6.1 (2011), pp. 1–23.

[89] Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. *Deep Reinforcement Learning for Sepsis Treatment*. 2017. arXiv: 1711.09602 [cs.AI].

[90] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. "Continuous State-Space Models for Optimal Sepsis Treatment: A Deep Reinforcement Learning Approach". In: *Machine Learning for Healthcare Conference*. PMLR. 2017, pp. 147–163.

[91] Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. *Model-Based Reinforcement Learning for Sepsis Treatment*. 2018. arXiv: 1811.09602 [cs.LG].

[92] James M Rehg, Susan A Murphy, and Santosh Kumar. "Mobile Health". In: *Cham: Springer International Publishing* (2017).

[93] James M Robins. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect". In: *Mathematical Modelling* 7.9-12 (1986), pp. 1393–1512.

[94] James M Robins. "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models". In: *Communications in Statistics-Theory and methods* 23.8 (1994), pp. 2379–2412.

[95] James M Robins. "Correction for Non-Compliance in Equivalence Trials". In: *Statistics in Medicine* 17.3 (1998), pp. 269–302.

[96] James M Robins. "Estimation of the Time-Dependent Accelerated Failure Time Model in the Presence of Confounding Factors". In: *Biometrika* 79.2 (1992), pp. 321–334.

[97] James M Robins. "Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference". In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Springer, 2000, pp. 95–133.

[98] James M Robins. "Optimal Structural Nested Models for Optimal Sequential Decisions". In: *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*. Springer. 2004, pp. 189–326.

[99] James M Robins. "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies". In: *Health Service Research Methodology: a Focus on AIDS* (1989), pp. 113–159.

[100] James M Robins, Liliana Orellana, and Andrea Rotnitzky. "Estimation and Extrapolation of Optimal Treatment and Testing Strategies". In: *Statistics in Medicine* 27.23 (2008), pp. 4678–4721.

[101] Luca Roggeveen, Ali El Hassouni, Jonas Ahrendt, Tingjie Guo, Lucas Fleuren, Patrick Thoral, Armand RJ Girbes, Mark Hoogendoorn, and Paul WG Elbers. "Transatlantic Transferability of a New Reinforcement Learning Model for Optimizing Hemodynamic Treatment for Critically Ill Patients with Sepsis". In: *Artificial Intelligence in Medicine* 112 (2021), p. 102003.

[102] Donald B Rubin. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment". In: *Journal of the American Statistical Association* 75.371 (1980), pp. 591–593.

[103] Phillip J Schulte, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. "Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes". In: *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 29.4 (2014), p. 640.

[104] Syed Ihtesham Hussain Shah, Antonio Coronato, and Muddasar Naeem. "Inverse Reinforcement Learning Based Approach for Investigating Optimal Dynamic Treatment Regime". In: *Workshops at 18th International Conference on Intelligent Environments (IE2022)*. IOS Press. 2022, pp. 266–276.

[105] Daniel Shin, Daniel S. Brown, and Anca D. Dragan. *Offline Preference-Based Apprenticeship Learning*. 2022. arXiv: 2107.09251 [cs.LG].

[106] Chamani Shiranthika, Kuo-Wei Chen, Chung-Yih Wang, Chan-Yun Yang, BH Sudantha, and Wei-Fu Li. "Supervised Optimal Chemotherapy Regimen Based on Offline Reinforcement Learning". In: *IEEE Journal of Biomedical and Health Informatics* 26.9 (2022), pp. 4763–4772.

[107] Susan M Shortreed, Eric Laber, Daniel J Lizotte, T Scott Stroup, Joelle Pineau, and Susan A Murphy. "Informing Sequential Clinical Decision-Making Through Reinforcement Learning: An Empirical Study". In: *Machine Learning* 84.1-2 (2011), p. 109.

[108] Moshe Sniedovich. "A New Look at Bellman's Principle of Optimality". In: *Journal of Optimization Theory and Applications* 49 (1986), pp. 161–176.

[109] Aaron Sonabend, Junwei Lu, Leo Anthony Celi, Tianxi Cai, and Peter Szolovits. "Expert-Supervised Reinforcement Learning for Offline Policy Learning and Evaluation". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 18967–18977.

[110] Rui Song, Michael Kosorok, Donglin Zeng, Yingqi Zhao, Eric Laber, and Ming Yuan. "On Sparse Representation for Optimal Individualized Treatment Selection with Penalized Outcome Weighted Learning". In: *Stat* 4.1 (2015), pp. 59–68.

[111] John Sperger, Nikki LB Freeman, Xiaotong Jiang, David Bang, Daniel de Marchi, and Michael R Kosorok. "The Future of Precision Health Is Data-Driven Decision Support". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 13.6 (2020), pp. 537–543.

[112] Richard S Sutton. "Planning by incremental dynamic programming". In: *Machine learning proceedings 1991*. Elsevier, 1991, pp. 353–357.

[113] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[114] Shengpu Tang, Aditya Modi, Michael Sjoding, and Jenna Wiens. "Clinician-in-the-Loop Decision Making: Reinforcement Learning with Near-Optimal Set-Valued Policies". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9387–9396.

[115] Sharon F Terry. "Obama's Precision Medicine Initiative". In: *Genetic testing and Molecular Biomarkers* 19.3 (2015), p. 113.

[116] Peter F Thall, Hsi-Guang Sung, and Elihu H Estey. "Selecting Therapeutic Strategies Based on Efficacy and Death in Multicourse Clinical Trials". In: *Journal of the American Statistical Association* 97.457 (2002), pp. 29–39.

[117] Peter F Thall, Leiko H Wooten, Christopher J Logothetis, Randall E Millikan, and Nizar M Tannir. "Bayesian and Frequentist Two-Stage Treatment Strategies Based on Sequential Failure Times Subject to Interval Censoring". In: *Statistics in Medicine* 26.26 (2007), pp. 4687–4702.

[118] Huan-Hsin Tseng, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K Ten Haken, and Issam El Naqa. "Deep Reinforcement Learning for Automated Radiation Adaptation in Lung Cancer". In: *Medical Physics* 44.12 (2017), pp. 6690–6705.

[119] Anastasios A Tsiatis, Marie Davidian, Shannon T Holloway, and Eric B Laber. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman and Hall/CRC, 2019.

[120] Alkeos Tsokos, Santhosh Narayanan, Ioannis Kosmidis, Gianluca Baio, Mihai Cucuringu, Gavin Whitaker, and Franz Király. "Modeling Outcomes of Soccer Matches". In: *Machine Learning* 108 (2019), pp. 77–95.

[121] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. *A Review of Off-Policy Evaluation in Reinforcement Learning*. 2022. arXiv: 2212.06355 [stat.ML].

[122] Cristiano Varin and David Firth. *Ridge Regression for Paired Comparisons: A Tractable New Approach, with Application to Premier League Football*. 2024. arXiv: 2406.09597 [stat.ME].

[123] Robert Vincent. "Reinforcement learning in models of adaptive medical treatment strategies". PhD thesis. 2014.

[124] Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. "Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic reatment Recommendation". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 2447–2456.

[125] Christopher JCH Watkins and Peter Dayan. "Q-Learning". In: *Machine Learning* 8 (1992), pp. 279–292.

[126] Christopher John Cornish Watkins. "Learning from Delayed Rewards". PhD thesis. King's College, 1989.

[127] Justin Weltz, Alex Volfovsky, and Eric B Laber. "Reinforcement Learning Methods in Public Health". In: *Clinical Therapeutics* 44.1 (2022), pp. 139–154.

[128] XiaoDan Wu, RuiChang Li, Zhen He, TianZhi Yu, and ChangQing Cheng. "A Value-Based Deep Reinforcement Learning Model with Human Expertise in Optimal Treatment of Sepsis". In: *npj Digital Medicine* 6.1 (2023), p. 15.

[129] Yanxun Xu, Peter Müller, Abdus S Wahed, and Peter F Thall. "Bayesian Nonparametric Estimation for Dynamic Treatment Regimes with Sequential Transition Times". In: *Journal of the American Statistical Association* 111.515 (2016), pp. 921–950.

[130] Gregory Yauney and Pratik Shah. "Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection". In: *Machine Learning for Healthcare Conference*. PMLR. 2018, pp. 161–226.

[131] Parisa Yazdjerdi, Nader Meskin, Mohammad Al-Naemi, Ala-Eddin Al Moustafa, and Levente Kovács. "Reinforcement Learning-Based Control of Tumor Growth Under Anti-Angiogenic Therapy". In: *Computer Methods and Programs in Biomedicine* 173 (2019), pp. 15–26.

[132] Chao Yu, Yinzhao Dong, Jiming Liu, and Guoqi Ren. "Incorporating Causal Factors into Reinforcement Learning for Dynamic Treatment Regimes in HIV". In: *BMC Medical Informatics and Decision Making* 19 (2019), pp. 19–29.

[133] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. "Reinforcement Learning in Healthcare: A Survey". In: *ACM Computing Surveys (CSUR)* 55.1 (2021), pp. 1–36.

[134] Chao Yu, Guoqi Ren, and Yinzhao Dong. "Supervised-Actor-Critic Reinforcement Learning for Intelligent Mechanical Ventilation and Sedative Dosing in Intensive Care Units". In: *BMC Medical Informatics and Decision Making* 20.3 (2020), pp. 1–8.

[135] Chao Yu, Guoqi Ren, and Jiming Liu. "Deep Inverse Reinforcement Learning for Sepsis Treatment". In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. 2019, pp. 1–3.

[136] Weichang Yu and Howard D Bondell. "Bayesian Likelihood-Based Regression for Estimation of Optimal Dynamic Treatment Regimes". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85.3 (2023), pp. 551–574.

[137] Amir Ebrahimi Zade, Seyedhamidreza Shahabi Haghighi, and Madjid Soltani. "Reinforcement Learning for Optimal Scheduling of Glioblastoma Treatment with Temozolomide". In: *Computer Methods and Programs in Biomedicine* 193 (2020), p. 105443.

[138] Tristan Zajonc. "Bayesian Inference for Dynamic Treatment Regimes: Mobility, Equity, and Efficiency in Student Tracking". In: *Journal of the American Statistical Association* 107.497 (2012), pp. 80–92.

[139] Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. "Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions". In: *Biometrika* 100.3 (2013).

[140] Junzhe Zhang. "Designing Optimal Dynamic Treatment Regimes: A Causal Reinforcement Learning Approach". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 11012–11022.

[141] Ying-Qi Zhao, Eric B Laber, Yang Ning, Sumona Saha, and Bruce E Sands. "Efficient augmentation and relaxation learning for individualized treatment rules using observational data". In: *Journal of Machine Learning Research (JMLR)* 20 (2019).

[142] Ying-Qi Zhao, Donglin Zeng, Eric B Laber, and Michael R Kosorok. "New statistical learning methods for estimating optimal dynamic treatment regimes". In: *Journal of the American Statistical Association* 110.510 (2015), pp. 583–598.

[143] Ying-Qi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. "Estimating Individualized Treatment Rules Using Outcome Weighted Learning". In: *Journal of the American Statistical Association* 107.449 (2012), pp. 1106–1118.

[144] Yufan Zhao, Michael R Kosorok, and Donglin Zeng. "Reinforcement Learning Design for Cancer Clinical Trials". In: *Statistics in Medicine* 28.26 (2009), pp. 3294–3315.

[145] Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. "Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer". In: *Biometrics* 67.4 (2011), pp. 1422–1433.

[146] Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. "Residual Weighted Learning for Estimating Individualized Treatment Rules". In: *Journal of the American Statistical Association* 112.517 (2017), pp. 169–187.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Titre : Apprentissage par renforcement et outcome-weighted learning bayésien pour la médecine de précision. Intégration de connaissances médicales dans les algorithmes de décision.

Mots clés : Dynamic treatment regimes, Individualized TreaIndividualized treatment regimetment Regimes, Processus de décision, Savoir d'expert, Apprentissage par préferences, Quantification d'incertitude

Résumé : La médecine de précision vise à adapter les traitements aux caractéristiques de chaque patient en s'appuyant sur les formalismes des "Individualized Treatment Regimes" (ITR) et des "Dynamic Treatment Regimes" (DTR). Les ITR concernent une seule décision thérapeutique, tandis que les DTR permettent l'adaptation des traitements au fil du temps via une séquence de décisions. Pour être pertinentes, ces approches doivent être en mesure de traiter des données complexes et d'intégrer les connaissances médicales, essentielles pour permettre une utilisation clinique réaliste et sans risques. Cette thèse présente trois projets de recherche. Premièrement, un état de l'art des méthodes d'intégration des connaissances médicales dans les modèles de "Reinforcement Learning" (RL) a été réalisé, en tenant compte du contexte des DTR et de leurs contraintes spécifiques pour une application sur des données observationnelles. Deuxièmement, une méthode probabiliste de construction des récompenses a été développée pour les modèles de RL, s'appuyant sur les préférences des experts médicaux. Illustrée par des études de cas sur le diabète et le cancer, cette méthode génère des récompenses de manière à exploiter les données, le savoir de l'expert médical et les relations entre les patients, évitant les biais de construction "à la main" et garantissant une cohérence avec les objectifs médicaux. Troisièmement, un cadre bayésien pour la méthode "Outcome-Weighted Learning" (OWL) a été proposé afin de quantifier l'incertitude dans les recommandations de traitement, renforçant ainsi la robustesse des décisions thérapeutiques, et a été illustré à travers de simulations de données. Les contributions de cette thèse visent à améliorer la fiabilité des outils de prise de décision en médecine de précision, d'une part en intégrant les connaissances médicales dans les modèles de RL, et d'autre part en proposant un cadre bayésien pour quantifier l'incertitude dans le modèle OWL. Ces travaux s'inscrivent dans une perspective globale de collaboration interdisciplinaire en particulier entre les domaines de l'apprentissage automatique, des sciences médicales et des statistiques.

Title: Reinforcement learning and Bayesian outcome-weighted learning for precision medicine: integrating medical knowledge into decision-making algorithms.

Key words: Dynamic treatment regimes, Individualized treatment regime, Decision process, Expert knowledge, Preference learning, Uncertainty quantification

Abstract: Precision medicine aims to tailor treatments to the characteristics of each patient by relying on the frameworks of Individualized Treatment Regimes (ITR) and Dynamic Treatment Regimes (DTR). ITRs involve a single therapeutic decision, while DTRs allow for the adaptation of treatments over time through a sequence of decisions. For these approaches to be effective, they must be capable of handling complex data and integrating medical knowledge, which is essential for enabling realistic and safe clinical use. This work presents three research projects. First, a state-of-the-art review of methods for integrating medical knowledge into Reinforcement Learning (RL) models was conducted, considering the context of DTR and their specific constraints for application to observational data. Second, a probabilistic method for constructing rewards was developed for RL models, based on the preferences of medical experts. Illustrated by case studies on diabetes and cancer, this method generates data-driven rewards, avoiding the biases of "manual" construction and ensuring consistency with medical objectives in learning treatment recommendation strategies. Third, a Bayesian framework for the Outcome-Weighted Learning (OWL) method was proposed to quantify uncertainty in treatment recommendations, thereby enhancing the robustness of therapeutic decisions, and was illustrated through simulations studies. This contributions aim to improve the reliability of decision-making tools in precision medicine, by integrating medical knowledge into RL models on one hand, and proposing a Bayesian framework to quantify uncertainty in the OWL model on the other. This work is part of a global perspective of interdisciplinary collaboration, particularly among the fields of machine learning, medical sciences, and statistics.