ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

PARCOURS DATASCIENCE
Master Mathématiques et Applications

Année universitaire 2018-2019

Apprentissage par renforcement

The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care

M. Komorowski, LA. Celi, O. Badawi, AC. Gordon AA. Faisal. *Nature Medicine*. Nov 2018

Philippe Burgelin
Gauthier Perrod
Manuel Pichon

Introduction

Sur le papier, le message apparaît sans équivoque : la santé et ses applications sont un axe d'intégration prioritaire des techniques de machine learning (rapport *Stratégie France IA 2017*). Cependant, si les progrès sont majeurs dans certaines spécialités (Computer Vision en radiologie, Data Mining en génomique appliquée à la cancérologie), d'autres domaines de la médecine semblent réfractaires aux promesses de l'intelligence artificielle et du machine learning. Le milieu de la réanimation fait sans doute partie de ces villages qui résistent encore et toujours à l'envahisseur numérique. Ce qui semble paradoxal au premier abord quand on connaît l'étendue des moyens techniques à disposition des équipes de soins pour prendre en charge ces patients complexes, ainsi que les capacités d'exploitation de nombreuses données collectées plusieurs fois par jour au cours de la surveillance continue des patients.

D'autre part, si certains problèmes d'optimisation de stratégies thérapeutiques ne sont pas des candidats idéaux pour une résolution par des algorithmes classiques d'apprentissage supervisé ou non supervisé, l'apprentissage par renforcement a démontré ces dernières années des résultats performants dans la modélisation d'interactions complexes entre un agent et son environnement. Ainsi, souhaitant étendre son utilisation à de nouveaux domaines autres que les jeux vidéos ou les véhicules autonomes, quelques équipes ont commencé à travailler sur le sujet en réanimation, sur des problématiques bien définies.

C'est le cas de l'équipe anglo-américaine à l'origine de la publication **"The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care"**, paru dans la revue *Nature Medicine* en Novembre 2018. Dans cet article les auteurs (M. Komorowski, LA. Celi, O. Badawi, AC. Gordon AA. Faisal.) proposent l'implémentation d'un modèle d'apprentissage par renforcement pour aider à optimiser l'état cardiocirculatoire (= **hémodynamique**) des patients de réanimation souffrant d'infections sévères (= regroupées sous la terminologie de **sepsis**).

Brièvement, l'état de sepsis correspond à l'apparition au cours d'une infection d'une ou plusieurs défaillances d'organes, souvent complexes, suite à une réaction immunitaire disproportionnée de l'hôte infecté. La fréquence de cette pathologie dans le monde, en particulier chez des sujets fragilisés par d'autres comorbidités, ainsi que son taux élevé de mortalité en font toujours une priorité de santé publique, particulièrement en réanimation. Cependant, malgré la connaissance de plus en plus précise des mécanismes à l'oeuvre et des aspects multiples du traitement, il semble difficile pour les équipes médicales les plus expertes de réduire le taux de mortalité au-delà d'un certain seuil. Parmi les facteurs incriminés, la grande variabilité interindividuelle des profils évolutifs, ainsi que la variabilité temporelle pour un patient donné, jouent probablement un rôle important.

L'hypothèse formulée par les auteurs et de prouver l'apport de méthodes d'apprentissage par renforcement pour individualiser et optimiser une stratégie thérapeutique en environnement critique complexe avec une dynamique évolutive importante comme les états de sepsis. Ils se sont focalisés sur le traitement des complications hémodynamiques initiales du sepsis (= chute de la pression artérielle, donc mauvaise perfusion des tissus, chute de l'apport en oxygène, et donc souffrance cellulaire généralisée).

Notre travail dans ce rapport sera d'explorer leur méthodologie et de la rapprocher des concepts vus en cours, tout en développant certaines notions spécifiques.

Nous nous placerons pour la suite en contexte d'apprentissage par renforcement.

Ainsi :

- Le temps est considéré comme s'écoulant de manière discrète. Un instant est noté $t \in \mathbb{N}$
- La situation étudiée est celle d'un agent interagissant avec son environnement. À chaque instant t, il perçoit un état s_t, et réalise une action a_t. Suite à cela, il reçoit une récompense de son action r_t. Il transite alors vers un nouvel état s_{t+1}.
- On définit une politique (ou stratégie) π comme une application :

$$\begin{aligned} \pi : \quad & \mathcal{S} \times \mathcal{A} \to [0,1] \\ & (s,a) \mapsto \pi(s,a) = P[a_t = a | s_t = s] \end{aligned}$$

- On définit la récompense $\mathcal{R}_t$ reçue par l'agent à partir de l'instant t par :

$$\mathcal{R}_t = \sum_{k=0} \gamma^k \, r_{t+k}$$

- On définit la "fonction valeur" d'un état s pour une politique fixée π par :

$$V^\pi(s) = \mathbb{E}[\mathcal{R}_t | s_t = s]$$

- On définit la "fonction action-valeur" d'un couple (état, action) pour une politique fixée π par :

$$Q^\pi(s,a) = \mathbb{E}[\mathcal{R}_t | s_t = s, a_t = a]$$

- Objectif : trouver la politique que doit suivre l'agent pour maximiser $\mathcal{R}_0$. Cette politique est qualifiée d'optimale. On la note π^*.

Note : ayant eu l'opportunité d'être en contact depuis le mois de décembre avec l'auteur principal de la publication, nous avions espéré initialement proposer une implémentation en Python de la méthode décrite au travers de l'article, pour retrouver les résultats principaux de l'étude.

Malheureusement, après quelques échanges par mail avec l'auteur en question pour demander un partage du dataset préparé exploitable, celui-ci nous a confié avoir essuyé le refus de ses superviseurs au motif d'une crainte de concurrence entre équipes universitaires différentes.

La base de données MIMIC-III étant décrite comme "publique", et disposant des éléments de code en MATLAB pour le preprocessing utilisé par les auteurs, nous avons pris le parti d'essayer de répliquer un dataset approché à partir des données brutes. Il fallait pour cela valider au préalable un cours en ligne sur l'éthique et la réglementation encadrant la recherche clinique aux USA ("Data or Specimens Only Research", CITI program).

Accédant finalement à la base de données souhaitée, nous avons été confrontés à de trop volumineuses données (> 30 Go pour certains fichiers .csv), avec des index particuliers, ce qui aurait impliqué beaucoup de temps de preprocessing. Nous avons donc choisi en dernier recours d'abandonner l'approche initiale pour

une approche plus théorique à défaut de pouvoir réaliser une implémentation correcte de la méthodologie présentée.

Nous proposerons cependant en *Annexes* un exemple d'implémentation envisageable en pseudo-code.

Contents

1 Méthodologie utilisée **5**

1.1 Données . . . 5
1.2 Critères d'inclusion . . . 5
1.3 Extraction des données et preprocessing . . . 6
1.4 Construction du modèle : Processus de Décision de Markov . . . 7
1.4.1 Espace d'états $\mathcal{S}$ . . . 7
1.4.2 Espace d'actions $\mathcal{A}$ . . . 8
1.4.3 Matrice de transition $\mathcal{T}(s', s, a)$ . . . 8
1.4.4 Matrice de récompense $\mathcal{R}(s', s, a)$ . . . 9
1.4.5 Coefficient de dépréciation γ . . . 9
1.5 Evaluation des actions des cliniciens par Q-learning . . . 11

2 Estimation et évaluation de la politique AI **12**

2.1 Estimation de la politique AI . . . 12
2.2 Evaluation de la politique AI . . . 12
2.2.1 HCOPE : High Confidence Off-Policy Evaluation . . . 13
2.2.2 Weighted Importance Sampling et Bootstraping . . . 14

3 Résultats et Discussion **15**

4 Conclusion **16**

5 Annexes **18**

5.1 Exemple d'implémentation . . . 18

Méthodologie utilisée

Nous allons détailler dans cette partie les étapes successives de la méthodologie développée par les auteurs, en nous efforçant d'expliciter au maximum les éléments de modélisation issus des concepts étudiés dans le cours.

Afin de faciliter le repérage du lecteur nous proposons en Figure 1.2 un flowchart synthétique reprenant la chronologie de ces différentes étapes.

1.1 Données

Le développement de l'outil *"AI Clinician"* (nous conserverons par la suite cette dénomination) s'est appuyé sur l'exploitation de **deux bases de données massives** exhaustives, distinctes, contenant l'ensemble des informations cliniques et paracliniques anonymisées issues du monitorage de plus de 80,000 patients de réanimation nord-américains : la base de données **MIMIC-III** (Medical Information Mart for Intensive Care version III) et la base **eRI** (eICU Research Institute Database), utilisées respectivement pour l'apprentissage du modèle puis sa validation.

Les données contenues dans ces deux registres définissent deux échantillons de population globalement comparables, dont les principales caractéristiques sont résumées dans le Tableau 5.1 en *Annexes*.

1.2 Critères d'inclusion

Le critère d'inclusion retenu pour la sélection des patients était la **survenue d'un sepsis** diagnostiqué à partir des critères actualisés de la dernière conférence de consensus internationale SEPSIS-3, à savoir une ou plusieurs dysfonction(s) d'organe menaçant le pronostic vital (évaluée par le score de gravité SOFA) et causée(s) par une réponse inappropriée de l'hôte à une infection suspectée ou documentée. Le plus précoce des événements entre l'administration d'un traitement anti-infectieux et la réalisation de prélèvements microbiologiques permettait de définir le début de l'épisode pathologique.

Il existait quelques critères d'exclusion, habituels ou justifiés par le contenu des bases de données :

- **Dans les deux registres MIMIC-III et eRI** :
  - Patients mineurs à la date d'admission
  - Mortalité non documentée
  - Limitation ou arrêt des thérapeutiques actives
- **Dans MIMIC-III**
  - Quantité de solutés administrés non documentée
- **Dans eRI**
  - Présence de réadmission (risque de mélange de données pour un même patient)
  - Données insuffisantes

1.3 Extraction des données et preprocessing

Afin de se focaliser sur l'optimisation du management hémodynamique du sepsis pendant les premières heures d'évolution, comprenant notamment la phase initiale de stabilisation riche en actes thérapeutiques potentiellement délétères, les données sélectionnées consistaient pour chaque variable d'intérêt en une **série temporelle discrétisée de 72 heures** d'évolution encadrant la date du diagnostic initial (classiquement, de 24h avant à 48h après). La **période entre deux valeurs successives** d'une même variable **était de 4h**. Il pouvait s'agir de la valeur réelle de recueil ou d'une moyenne pour des variables à périodicité plus courte.

Pour chaque patient sélectionné, les auteurs ont choisi de retenir **48 variables d'intérêt** permettant d'offrir une description vectorielle de l'état clinique du patient la plus représentative possible, bien que nécessairement partielle. Il s'agissait ainsi principalement de variables :

- à valeur démographique
- descriptives de la gravité de son évolution
- cliniques (fréquence cardiaque, pression artérielle, saturation, température, etc.)
- paracliniques (principalement des résultats d'examens de laboratoire décrivant les paramètres métaboliques du patient et les résultats microbiologiques)
- relatives aux thérapeutiques administrées
- de survie

Le descriptif de ces variables est présenté dans le Tableau 5.1 en *Annexes*. L'explication détaillée du rôle de chacune d'entre elles n'est pas indispensable pour la suite. En revanche nous reviendrons dans la fin de ce texte sur l'importance relative de ces variables pour éclairer la comparaison des performances décisionnelles du modèle algorithmique par rapport à celles des médecins.

La présence d'erreurs et de valeurs aberrantes a été contrôlée (méthode statistique indiquée par les auteurs : méthode de Tukey) et les corrections nécessaires ont été appliquées (transformations d'unités, restrictions à une borne de variation pour respecter une plausibilité clinique). Enfin deux dernières opérations ont permis de finaliser la préparation des données avant de débuter la construction du modèle :

- **la complétion de données manquantes**, indispensable compte tenu de la nécessité de disposer d'un jeu de données complet pour appliquer l'algorithme de clustering (k-means) permettant de définir l'espace d'états du problème. Les données ont été ainsi soit interpolées à partir de la dernière valeur disponible quand cela était pertinent, soit en utilisant la méthode des k plus proches voisins (k-NN);
- **la standardisation** adaptée à la distribution statistique des données pour s'affranchir de l'hétérogénéité des différentes variables.

A partir des échantillons préparés issus des registres MIMIC-III et eRI, **trois datasets** ont été constitués :

- **une base de données d'apprentissage** d'un volume équivalent à 80% du registre MIMIC-III, pour la construction des modèles d'apprentissage par renforcement;
- **une base de données de validation intermédiaire** d'un volume équivalent à 20% du regsitre MIMIC-III, pour sélectionner le modèle optimal parmi 500 versions différentes du modèle construit à partir de 80% de MIMIC-III;
- **une base de données de validation définitive** issue de la préparation du registre eRI, pour tester les performances du meilleur modèle.

1.4 Construction du modèle : Processus de Décision de Markov

L'hypothèse de modélisation principale des auteurs est de représenter l'évolution de l'état clinique d'un patient, partiellement représenté par l'ensemble des paramètres d'intérêt sélectionnés, par un **Processus de Décision de Markov**. La propriété fondamentale et usuelle des environnements markoviens réside dans la dépendance unique à l'état antérieur pour les paramètres évolutifs.

Dans un modèle markovien stationnaire on a la probabilité de rester dans un même état qui décroît exponentiellement, en effet pour un modèle avec une mémoire d'uniquement une étape :

$$P(X, X, X, X..., X|X) = \prod_{t=1}^{T} P(X|X) = P(X|X)^T$$

Afin de valider leur hypothèse de modélisation, les auteurs ont donc vérifié empiriquement au préalable cette propriété, comme illustré sur la Figure 1.1.

Dans le contexte spécifique de l'apprentissage par renforcement, la définition d'un agent suivant un Processus de Décision de Markov nécessite de connaître le quadruplet de paramètres suivants :

- $\mathcal{S}$ un **ensemble d'états fini** (= l'ensemble des états cliniques envisageables pour des patients de réanimation atteints de sepsis, dans notre cas d'étude);
- $\mathcal{A}$ un **ensemble d'actions fini** (= l'ensemble des options thérapeutiques envisagées pour améliorer l'état clinique du patient, dans notre cas d'étude). $\mathcal{A}(s)$ est le sous-ensemble des actions possibles à l'état s;
- $\mathcal{T}(s', s, a)$ la **matrice stochastique de transition**, permettant de décrire complètement la dynamique de l'environnement. Chaque élément de cette matrice représente la probabilité $p(s'|s, a)$ d'obtenir l'état $s' \in \mathcal{S}$ au temps $t + 1$ connaissant l'état $s \in \mathcal{S}$ au temps t et l'action $a \in \mathcal{A}$ choisie au même temps t.
- $\mathcal{R}(s', s, a)$ la **matrice de récompense** où chaque élément représente l'espérance de la récompense immédiate reçue par l'agent conditionnellement à la transition de l'état s vers l'état s' via l'action a (= dans notre cas, récompense si la transition s'inscrit dans une trajectoire synonyme de survie du patient, pénalité dans le cas contraire).

1.4.1 Espace d'états $\mathcal{S}$

Les différents états cliniques possibles pour chacun des patients septiques sélectionnés ont été obtenus après utilisation d'un **algorithme de clustering de type "k-means"** sur la base de données d'apprentissage (= 80% de MIMIC-III). Le nombre optimal de classes compte tenu du nombre de données disponibles - $n = 750$ dans l'article - a été déterminé par comparaison des critères d'information AIC et BIC entre les différents modèles de clustering.

L'objectif était d'obtenir un ensemble d'états suffisamment granulaire pour représenter le plus fidèlement possible le continuum d'évolution observé en pratique clinique.

Il fallait évidemment éviter un nombre de classes trop élevé - typiquement supérieur à 1000 - synonyme de représentation trop parcimonieuse de chaque classe.

A noter que **deux états supplémentaires dits absorbants** ont été rajoutés, correspondant chacun au décès du patient ou à sa sortie de l'unité au cours de la période de suivi de 72h.

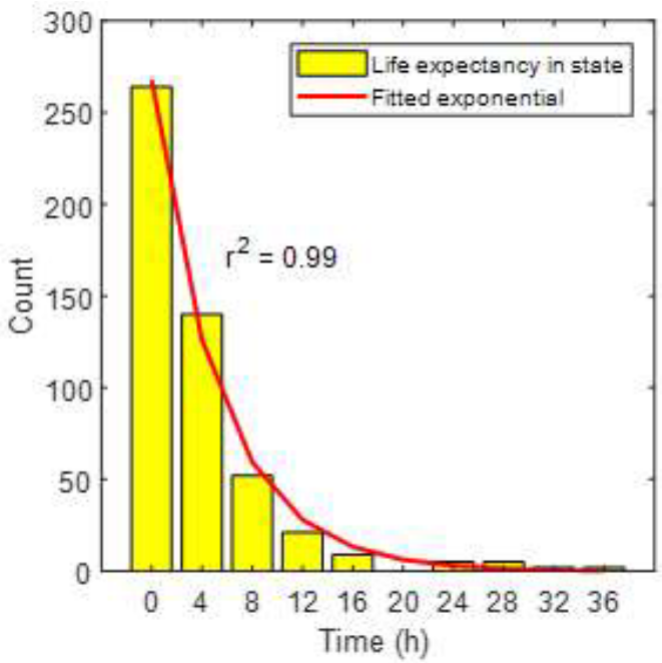

Figure 1.1 – Vérification d'une propriété markovienne : décroissance exponentielle pour la probabilité de rester dans un état donné

1.4.2 Espace d'actions $\mathcal{A}$

La prise en charge des infections sévères en réanimation relevant de décisions thérapeutiques complexes agissant sur plusieurs dimensions (traitement anti-infectieux précoce, restauration de l'état hémodynamique et souvent respiratoire, gestion des défaillances d'organes secondaires, etc.) et visant à pallier les défaillances d'organes multiples et imbriquées, les auteurs ont proposé par choix de simplification de ne s'intéresser qu'à l'optimisation de la défaillance cardio-circulatoire, systématique dans les états septiques graves. En apparence restrictif, cet aspect du traitement constitue cependant un des piliers de la stratégie thérapeutique globale, et son efficacité - en particulier dans les premières heures d'évolution de la pathologie - conditionne fortement la survenue de complications ultérieures. Ce choix semblait donc tout à fait pertinent.

Pour modéliser les interventions thérapeutiques à visée hémodynamique, destinées à assurer une pression de perfusion des tissus et un transport d'oxygène satisfaisants, les auteurs ont retenu les deux traitements les plus consensuels, à savoir **l'administration de solutés de remplissage vasculaire par voie intraveineuse** (= expansion volémique, pour augmenter le contenu liquidien à l'intérieur des vaisseaux) et **l'administration de traitements vasopresseurs** (= dont l'effet vasoconstricteur permet de diminuer le calibre des artères, et d'augmenter ainsi la pression de perfusion des organes).

Ils ont alors procédé à une **discrétisation** de l'ensemble des doses observées pour chacun de ces deux traitements, en **5 catégories de doses** allant de la catégorie "absence de traitement" à la catégorie "doses maximales administrées". En combinant les deux traitements possibles, on obtenait un **espace d'actions de cardinal 25**, chacune des classes définies étant représentée par la **dose médiane** à l'intérieur de la classe en question.

1.4.3 Matrice de transition $\mathcal{T}(s', s, a)$

La définition des espaces d'états et d'actions a permis de définir l'ensemble des **trajectoires** observées pour chaque patient présent dans la base de données d'apprentissage.

Pour obtenir la matrice de transition à partir de ces trajectoires, les auteurs ont ensuite procédé par **dénombrement** pour compter le **nombre d'occurrences de chaque transition entre deux états donnés**. En rapportant ce compte au nombre de transitions possibles à partir de chaque état, ils ont pu estimer les probabilités de transition et construire la matrice stochastique de transition.

En étudiant les actions observées sur l'ensemble des trajectoires, ils ont également pu restreindre la dimension du sous-espace $\mathcal{A}(s)$ des actions possibles à partir de l'état s aux actions les plus fréquemment observées (*i.e.* de cardinal > 5). Cette opération offre un **intérêt double** : **réduire l'espace des actions** aux éventualités les plus probables, et **assurer une certaine sécurité** en reproduisant les pratiques médicales observées qui évitent la prescription de traitements jugés absurdes (par exemple arrêter le traitement vasopresseur chez un patient nécessitant de fortes doses pour maintenir une hémodynamique satisfaisante).

1.4.4 Matrice de récompense $\mathcal{R}(s', s, a)$

Afin de comprendre le système de récompense mis en place par les auteurs pour permettre l'optimisation de la stratégie thérapeutique proposée par l'agent, il est nécessaire de revenir à l'objectif principal dans notre cadre d'étude, à savoir **maximiser la survie** des patients de réanimation atteints d'infection sévère en optimisant leur prise en charge hémodynamique pendant les 72 premières heures d'évolution.

La variable de décision initialement retenue était donc la **mortalité à 90 jours** (indicateur classique de survie dans les études cliniques en réanimation), variable non observée pendant la phase initiale ciblée pour le développement du modèle, car évaluée *a posteriori* (en dehors de la survenue malheureuse d'un décès pendant les 3 premiers jours d'évolution). La récompense décrite dans l'étude valant $+100$ en cas de survie à 90 jours, et -100 (pénalité) en cas de décès, elle peut être considérée comme un label des trajectoires observées en fonction de leur issue.

Ainsi de manière analogue à la méthode utilisée pour construire la matrice de transition, en procédant par **dénombrement** il est possible de compter pour chaque transition $(s, a) \to s'$ le nombre de fois où elle est impliquée dans une trajectoire "de survie" ou "de décès" et d'estimer ainsi l'espérance de la récompense immédiate associée à la transition de l'état s vers l'état s' via l'action a.

1.4.5 Coefficient de dépréciation γ

Facteur permettant de définir l'horizon de l'agent, il permet de régler l'importance relative attribuée aux récompenses futures par rapport aux récompenses immédiates. Fixé par les auteurs à 0.99, il incarne la volonté de développer un modèle optimisant sa politique pour maximiser la **survie à long terme** plutôt qu'à court terme uniquement. Sous réserve de validation des performances d'un tel outil, il s'agit là d'une perspective d'innovation majeure dans le management de pathologies réanimatoires aiguës, dans la mesure où il est très difficile pour un clinicien seul ou en équipe de prévoir l'impact de ses décisions à moyen ou à long terme.

Malgré la valeur élevée du facteur γ la convergence du modèle développé devrait pouvoir être assurée par le nombre fini d'états et d'actions.

Le modèle de Markov associé au futur agent étant désormais bien défini par ses paramètres de construction, nous pouvons préciser les méthodes d'estimation et d'évaluation des politiques *"médecin"* et *"AI Clinician"*.

Comme décrit sur le schéma de synthèse consultable sur la prochaine page, la suite de la méthodologie consiste à **générer 500 modèles différents** à partir d'un clustering à 750 classes opéré sur 500 échantillons aléatoires différents de taille "80% de MIMIC-III", puis **sélectionner le meilleur modèle avant validation finale sur la base de données de test eRI**, par une méthode particulière d'évaluation "off-policy" nommée **HCOPE** (pour high-confidence off-policy evaluation), basée sur une méthode de Weighted Importance Sampling, et développée dans la partie 2. de ce texte.

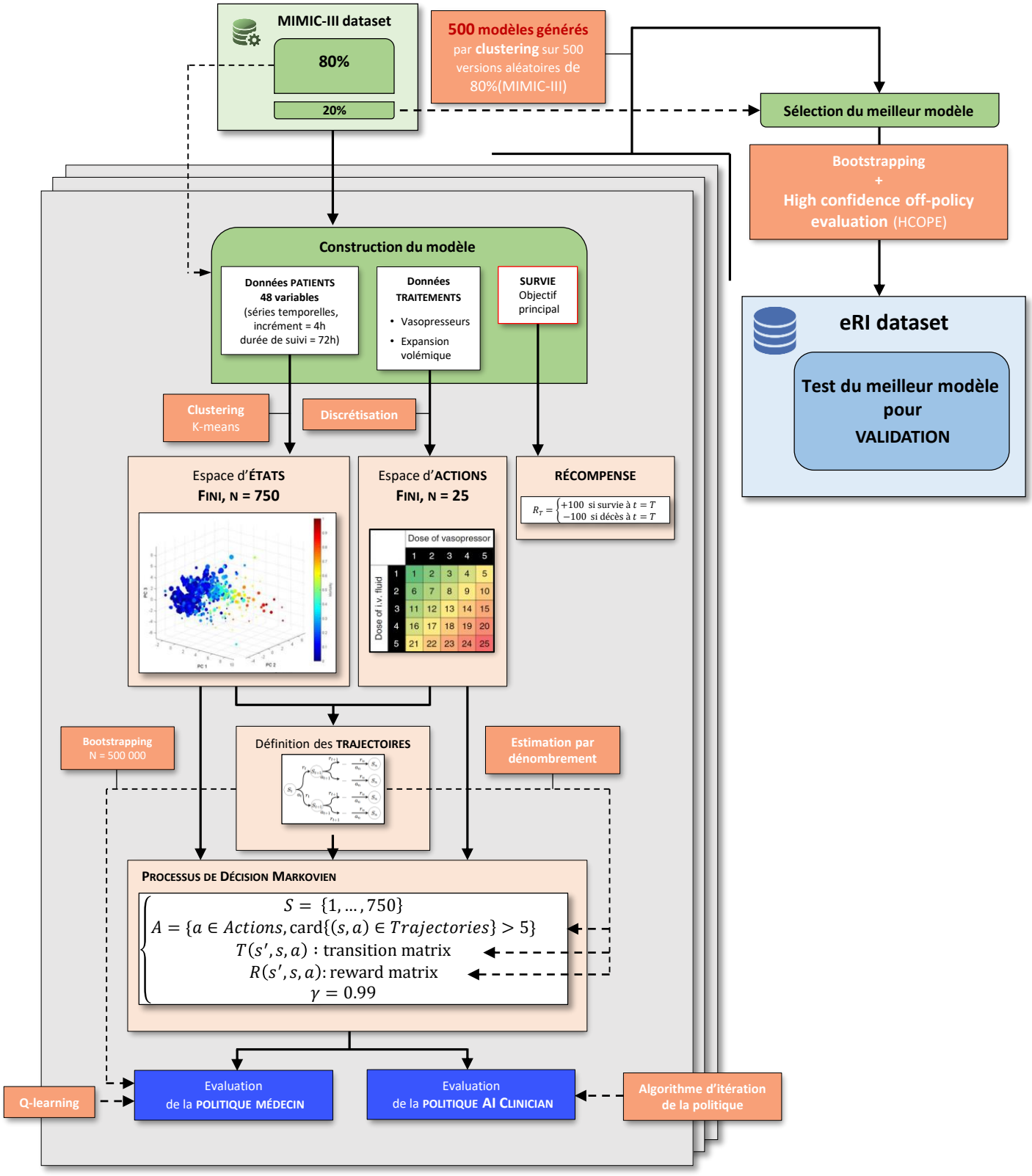

Figure 1.2 – Schéma résumé de la méthodologie développée

1.5 Evaluation des actions des cliniciens par Q-learning

Pour évaluer la pratique des cliniciens, il semble trop complexe d'essayer de construire un modèle d'actions-états complet. En revanche, il est très naturel et facile d'utiliser une méthode de **Q-learning**, c'est-à-dire d'évaluer la valeur attribuée à chaque action à partir de chaque état pour une politique donnée, car c'est ce que l'on observe dans notre monde et dans ce contexte. La méthode employée par les auteurs est une méthode dite de **off-policy TD-learning** (apprentissage par différence temporelle) d'ordre 0. Cette méthode est intéressante car elle offre un compromis entre une méthode d'exploration en profondeur (Monte-Carlo) ne nécessitant pas de modèle et une méthode qui nécessite un modèle (ex: par itérations). Il est alors possible d'apprendre seulement avec les données correspondant à la trajectoire de chaque patient.

On peut le décrire par l'algorithme suivant présenté dans [Sutton and Barto(2018)], la condition de convergence (= chaque paire *(état, action)* visité une infinité de fois) étant assurée dans l'étude par un rééchantillonnage par bootstrap de 500,000 nouvelles trajectoires patient :

Algorithm 1 Off-policy TD-learning

input: Taux d'apprentissage $\alpha \in (0, 1]$, ϵ petit
output: Q

Initialiser pour tout $s \in S$ et $a \in \mathcal{A}(s)$, $Q(s, a)$ de manière arbitraire sauf $Q(terminal, .) = 0$
Boucler pour partie (patient)
Initialiser S
Boucler pour chaque étape (observation)
Choisir une action A de S à partir d'une politique dérivée de Q (*)
Effectuer A, observer $S^{'}$ et R
$Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S^{'}, a) - Q(S, A)]$
$S \leftarrow S^{'}$
Jusqu'à ce que S soit terminal
Boucler
$\Delta \leftarrow 0$

(*) Une politique dérivée de Q peut consister en une politique ϵ-greedy dérivée de : $\pi(s) = \text{argmax}_a Q(s, a), \forall s$

Ainsi, on va pouvoir utiliser cet algorithme dans les 500 manières de créer les 750 clusters (donc les 500 espaces d'états). On va donc pouvoir établir 500 politiques différentes pour décrire la politique des cliniciens. Ces politiques seront utiles par la suite pour mettre en oeuvre du bootstrapping.

L'estimation et l'évaluation de la politique *"AI Clinician"*, ou *politique AI*, propose quelques particularités conceptuelles qui sont détaillées dans la prochaine partie.

Estimation et évaluation de la politique AI

2.1 Estimation de la politique AI

Une méthode classique pour construire une politique optimale est une itération in-place dans le but de maximiser la somme des récompenses sur le long terme.

Cette méthode est décrite dans [Sutton and Barto(2018)], dans l'algorithme suivant :

Algorithm 2 **Evaluation itérative de politique** L'input est une politique à évaluer π_e, un dataset de trajectoires, D, un niveau de confiance $\delta \in [0, 1]$ et un nombre d'estimation bootstrap B

---

input: Initialisation aléatoire de $V(s) \in R$ et $\pi(s) \in \mathrm{A}(s)$ arbitrairement $\forall s$
output: $V = v_*$ et $\pi = \pi_*$

Evaluation de la politique
Boucler
  $\Delta \leftarrow 0$
  Boucler sur chaque $s \in \mathrm{S}$
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_{s',r} p(s^{'}, r|s, \pi(s))[r + \gamma V(s^{'})]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
jusqu'à ce que $\Delta <$ Tolérance

Amélioration de la politique
politique-stable $\leftarrow$ *vrai*
Pour tout $s \in \mathrm{S}$
  ancienne-action $\leftarrow \pi(s)$
  $\pi(s) \leftarrow \mathrm{argmax}_a \sum_{s',r} p(s^{'}, r|s, \pi(s))[r + \gamma V(s^{'})]$
  Si *ancienne-action* $\neq \pi(s)$ alors *politique-stable* $\leftarrow faux$
Si *politique-stable* alors STOP

L'auteur ne précise pas dans l'article s'il a utilisé une méthode asynchrone ou synchrone (améliorer une partie des valeurs à chaque fois ou toutes les valeurs). Dans le cas où il aurait utilisé une méthode synchrone, une amélioration envisageable est d'utiliser une méthode asynchrone en améliorant par exemple en priorité les états qui nous semblent anormaux grâce à une connaissance médicale (mais nécessite un premier entraînement de l'algorithme et de détecter des anomalies).

2.2 Evaluation de la politique AI

Ayant accès à un ensemble de décisions thérapeutiques réelles de médecins dans l'environnement d'évolution de l'agent à optimiser, une méthode d'évaluation usuelle de la politique AI est de comparer la politique AI à la politique utilisée en pratique (= ici, la politique des cliniciens).

La méthode d'estimation et d'évaluation de la politique AI est donc une méthode dite **off-policy**, qui va utiliser des méthodes d'**Importance Sampling**, méthode classique en simulation de Monte-Carlo permettant de simuler une loi en ne connaissant qu'un majorant du ratio entre une loi cible et une loi connue selon laquelle on sait simuler (ratio inférieur à 1 et le plus proche possible de 1 pour être plus efficace en termes d'acceptation). On pourra néanmoins remarquer que dans l'article il est proposé d'adoucir la loi cible (ici,

elle sera 1% ϵ-greedy), à savoir la *politique AI*, et de ne pas utiliser une politique adoucie pour les médecins.

On pourrait alors avoir des problèmes lors du calcul du ratio $\frac{\pi_{cible}}{\pi_{clinicien}}$ car le numérateur ne vaut jamais 0 tandis que le dénominateur peut potentiellement valoir 0. D'un point de vue médical on peut se dire que si une trajectoire n'est jamais explorée par un médecin, il est peut être préférable qu'en effet l'algorithme n'explore pas non plus ces trajectoires et donc exclure ces possibilités de l'AI semble sécuritaire. C'est ainsi que l'article justifie ce choix, en contraignant l'AI à ne choisir que parmi des actions fréquemment observées à partir d'un certain état (= vues au moins plus de 5 fois).

2.2.1 HCOPE : High Confidence Off-Policy Evaluation

La méthode HCOPE est une méthode d'évaluation off-policy développée dans [Philip S. Thomas(2015)]. Celle-ci permet de donner un **intervalle de confiance assez restreint** lors de l'évaluation off-policy d'une nouvelle politique. Dans un contexte médical, et contrairement à d'autres domaines comme celui des jeux vidéos, il est bien entendu inenvisageable de tester directement sur des patients les décisions de politiques non sécuritaires qui pourraient conduire à des complications gravissimes voire au décès du patient. Il semble donc primordial de pouvoir condenser au maximum l'intervalle de confiance et en particulier de pouvoir estimer la borne inférieure de celui-ci.

La méthode HCOPE est en fait l'utilisation de très nombreuses *behavior policies* pour estimer notre politique par une méthode off-policy.

On note $\mathcal{D} = \{(\tau_i, \theta_i)$, i $\in \{1..n\}, \tau_i$ trajectoires générées par les politiques de paramètres $\theta_i\}$

On estime alors $\rho_\theta := \mathbf{E}[R_\tau|\theta]$ par son estimateur non biaisé par importance-sampling :

$$\hat{\rho}(\theta, \tau, \theta_i) = R_\tau \frac{\prod_{t=1}^{T} \pi(a_t|s_t\theta)}{\prod_{t=1}^{T} \pi(a_t|s_t\theta_i)}$$

Toutes ces évaluations permettent ensuite de calculer une borne inférieure, en utilisant le théorème de concentration présenté dans [Philip S. Thomas(2015)] :

Théorème :

Soit $X_1, ..., X_n$ des variables aléatoires réelles indépendantes presque sûrement positives, telles que $\forall i, \mathbf{E}[X_i] \leq \mu$, et que pour un seuil c, on pose $Y_i = \min(X_i, c)$ alors avec une probabilité au moins $1 - \delta$ on a :

$$\mu \geq \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{c}{n} \frac{7n \ln(\frac{2}{\delta})}{3(n-1)} - \frac{c}{n} \sqrt{\frac{\ln(\frac{2}{\delta})}{n-1} \sum_{i,j=1}^{n} \left(\frac{Y_i - Y_j}{c}\right)^2}$$

On obtient ainsi une borne inférieure de confiance $1 - \delta$ pour μ en optimisant sur c.

Pour ce faire il faut découper $\mathcal{D}$ en $\mathcal{D}_{pre}$ et $\mathcal{D}_{post}$, avec c estimé sur $\mathcal{D}_{pre}$ et la borne inférieure calculée sur $\mathcal{D}_{post}$.

Dans notre exemple médical, on vérifie bien toutes les hypothèses du théorème car les importance-weights sont forcéments positifs et toutes les espérances sont bornées par $+100$, la valeur maximale de récompense créditée si le patient survit.

Encore peu utilisée dans un tel contexte, la méthode HCOPE paraît ainsi offrir des garanties sécuritaires adaptées au problème présenté, car il est très important de pouvoir assurer une borne inférieure avec une forte probabilité dans le contexte critique de la survie d'un patient au pronostic vital engagé. Néanmoins la méthode est très coûteuse en données car un terme décroît à la vitesse $\sqrt{n}$. Pour améliorer ce résultat on utilise une technique de bootstrapping.

2.2.2 Weighted Importance Sampling et Bootstraping

Les méthodes de bootstrapping proposent des garanties théoriques solides lorsque le nombre d'observations tend vers l'infini. [Josiah P. Hanna and Niekum(2017)] se propose de présenter des résultats non asymptotiques sur le bootstrap en apprentissage par renforcement. Pour rappel, le bootstrapping est une technique consistant à utiliser de nombreuses fois ses observations en rééchantillonnant parmi celles-ci selon un procédé de tirage avec remise. Ici, les observations seront les trajectoires.

Une manière d'utiliser le bootstrap est présentée par l'algorithme suivant :

Algorithm 3 Intervalle de confiance en Bootstrap L'input est une politique à évaluer π_e, un dataset de trajectoires, D, un niveau de confiance $\delta \in [0, 1]$ et un nombre d'estimations bootstrap B

input: $\pi_e, \mathrm{D}, \pi_b, \delta, B$

output: L'intervalle de confiance sur la borne inférieure de la valeur de π_e.

1: **Pour tout** $i \in [1, B]$ **faire**
2: $\tilde{\mathrm{D}}_i \leftarrow \left\{H_1^i, .., H_n^i\right\}$ où $H_j^i \sim \mathrm{U(D)}$, avec U la distribution uniforme
3: $\hat{V}_i \leftarrow$ **Off-PolicyEstimate**$(\pi_e, \tilde{\mathrm{D}}_i, \pi_b)$ (#HCOPE)
4: **Terminer boucle**
5: Trier de manière croissante $\left\{\hat{V}_i | i \in [1, B]\right\}$
6: $l \leftarrow \lfloor \delta B \rfloor$
7: **Retouner** $\hat{V}_l$

Cet algorithme permet une estimation de la borne inférieure de la valeur de la politique. Dans [Josiah P. Hanna and Niekum(2017)], les exemples utilisent $B = 2000$ et c'est aussi cette valeur qui est utilisée dans l'article. Ce nombre semble cohérent avec les données que l'on a (500 politiques AI que l'on cherche à classer avec une off-policy estimation en utilisant les 500 politiques médecin).

Résultats et Discussion

Si l'objet de ce travail était surtout d'étudier la méthodologie développée pour concevoir un outil adapté d'apprentissage par renforcement, il est certainement intéressant de proposer quelques commentaires au sujet des résultats principaux de l'étude, à savoir la vraisemblable **capacité de l'outil "AI Clinician" à individualiser et optimiser une stratégie thérapeutique pour prendre en charge efficacement les désordres cardiocirculatoires associés à la phase initiale du sepsis, et ainsi diminuer la mortalité des patients concernés** :

- malgré la simplification inévitable du modèle, la configuration pertinente des paramètres dans un objectif de maximisation de survie à long terme est probablement un des facteurs-clés de la **bonne calibration du modèle**. En effet les traitements récompensés positivement étaient associés à une mortalité faible, et les traitements pénalisés à une mortalité plus élevée;

- l'évaluation par la **méthode HCOPE** offre probablement les **garanties suffisantes** concernant la consistance et les marges de sécurité des estimations statistiques pour envisager une étude prospective comparative testant en temps réel l'optimisation de stratégie thérapeutique par un algorithme de renforcement analogue;

- la **concordance** de certains éléments du modèle **avec des observations issues de plusieurs années de pratique clinique et de recherche médicale** est étonnante et très prometteuse, en particulier :

  - la mise en évidence de l'importance relative des paramètres décisionnels (classification par random forest) pour l'outil "AI Clinician" et pour les médecins montre une stratification très proche (voir Figure 5.2 en *Annexes*). Les décisions valorisées par l'IA sont donc cliniquement pertinentes et surtout élaborées principalement à partir de variables clinico-biologiques, comme le ferait une équipe de médecins en pratique clinique courante;
  - les stratégies thérapeutiques valorisées par l'IA corroborent l'évolution des résultats de la littérature scientifique sur le sepsis ces dernières années. Ainsi, alors qu'il ne fait aucun doute quant au rôle extrêmement délétère d'un remplissage vasculaire excessif en phase initiale lors d'un sepsis, il est très satisfaisant de constater que l'IA propose des doses modérées pour ce type de traitement. En revanche pour l'IA l'accent est mis sur l'utilisation plus précoce des traitements vasopresseurs, et cette recommandation va dans le sens des éléments fondamentaux de la physiopathologie du sepsis (vasoplégie majeure en phase initiale, corrigée par les traitements vasopresseurs) et également des dernières recommandations internationales;

- la méthodologie d'apprentissage par renforcement présentée dans cet article étant plutôt classique, le simple fait d'essayer de complexifier les actions possibles du modèle (pour peu que suffisamment de données soient disponibles) ou d'envisager des approches différentes (exemple de [Niranjani Prasad(2017)], article dans lequel est développée une méthode d'apprentissage par batch nécessitant de pouvoir appliquer une méthode de régression paramétrique (réseau de neurones, SVM, processus gaussiens...) afin d'optimiser la ventilation mécanique en réanimation) offre des perspectives d'évolution intéressantes, probablement à la hauteur des progrès spectaculaires accomplis ces dernières années dans la résolution de tâches complexes par les algorithmes d'apprentissage par renforcement.

Conclusion

Cette publication, qui reprend une méthodologie fréquemment rencontrée en apprentissage par renforcement, marque une innovation dans le domaine médical clinique.

Les résultats obtenus dans l'estimation de la réduction de mortalité par "simple" optimisation individualisée d'une stratégie simplifiée de prise en charge hémodynamique à partir de données exhaustives de patients septiques en réanimation sont très encourageants. Ils ouvrent la voie vers une médecine individualisée capable d'améliorer le pronostic des patients les plus lourds même en l'absence de révolution pharmaceutique.

La concordance des décisions suggérées par l'IA avec les éléments courants de la pratique clinique est manifeste. D'autre part, la conception d'un outil permettant de maximiser un objectif de survie à long terme (stratégie difficile à concevoir pour un médecin réanimateur conditionné par la recherche de bénéfices thérapeutiques à court terme) permettrait d'aiguiller les prescriptions des équipes médicales si l'efficacité d'un tel outil venait à être confirmée par des études ultérieures. Mais d'ores et déjà l'évaluation des performances algorithmiques par des méthodes sécuritaires de type HCOPE offrent des garanties solides sur le plan statistique, particulièrement adaptées au cadre éthique médical.

L'apprentissage par renforcement, par sa capacité à résoudre parfois "simplement" des problèmes décisionnels complexes pourrait donc vraisemblablement améliorer la pratique clinique notamment en réanimation compte tenu de la proportion importante de patients complexes dans ce type de service.

Bibliography

[INRIA(2014-2018)] INRIA, 2014-2018. Markov decision processes (mdp) toolbox.

[Josiah P. Hanna and Niekum(2017)] Josiah P. Hanna, P. S., Niekum, S., 2017. Bootstrapping with models: Confidence intervals for off-policy evaluation.

[Niranjani Prasad(2017)] Niranjani Prasad, e. a., 2017. A reinforcement learning approach to weaning of mechanical ventilation in intensive care unit.

[Philip S. Thomas(2015)] Philip S. Thomas, Georgios Theocharous, M. G., 2015. High confidence off-policy evaluation.

[Sutton and Barto(2018)] Sutton, R., Barto, A., 2018. Reinforcement Learning: An Introduction, second edition.

Annexes

5.1 Exemple d'implémentation

Comme annoncé en Introduction, nous proposons ci-après un exemple d'implémentation possible du modèle présenté dans l'article.

L'objectif est de présenter les grandes étapes du processus dans un langage logique. Il est important de noter que les matrices seront gérées ici comme en MATLAB.

La première étape du processus est l'algorithme de clustering de type "k-means".

```
1
2     ## k-means
3     for (modelToTest in 1:500)
4     {
5         k=750
6         center = createRandomCentroids(k, nbOfFeatures)
7         nbIter = 0
8
9         while (nbIter < nbIterMax) && (oldCenter != Center)
10        {
11            oldCenter = Center
12            nbIter++
13
14            for (i in 0:nbOfpatients)
15            {
16                cluster[i] = argmax(for j in 0:k) {euclidianDistance(x(i),
                      center[j])}
17            }
18            for (j in 0:k)
19            {
20                center[j] = mean(x[i] where cluster[i]=j)
21            }
22
23        }
24        clusterModel[modelToTest] = cluster
25        centerModel[modelToTest] = center
26    }
```

A partir des trajectoires identifées, il est possible d'estimer la matrice de transition.

```

## écration de matrice de transition

transition=zeros(ncl2,ncl2,nact);
sumsnclnact=zeros(ncl2,nact);

for (i in 1:size(qldata3,1)-1)
{
    if (data(i+1,1))~=1  ## if we are not in the last state for this patient = if there is a transition to make
    {
        S0=data(i,2)
        S1=data(i+1,2)
        acid= data(i,3)
        transition(S1,S0,acid)=transition(S1,S0,acid)+1
        sumsnclnact(S0,acid)=sumsnclnact(S0,acid)+1
    }

}

sumsnclnact(sumsnclnact<=transthres)=0;  ##delete rare transitions ( those seen less than 5 times = bottom 50% cf article)

 for (i in 1:ncl2)
     for (j in 1:nact)
     {
        if( sumsnclnact(i,j)==0)
        {
            transition(:,i,j)=0
        }
        else
        {
            transition(:,i,j)=transition(:,i,j)/sumsnclnact(i,j)
        }
     }
```

Puis création de la matrice de récompense.

```

## écration de matrice de ércompense (voir annexe de l'article pour le choix des variables)

recompense=zeros(ncl2,ncl2,nact) ##nact represente les actions
recompense[ncl1,:,:]=-100 ##output negatif quand la personne meurt
recompense[ncl2,:,:]=100 ##output positif quand la personne reste vivante
R=sum(transition.*recompense) ## transmition des ércompenses aux édiffrents états de transition
```

Enfin l'application de la off policy. On a utilisé ici une fonction développée dans la MDPToolbox disponible ici : http://www7.inra.fr/mia/T/MDPtoolbox/ [INRIA(2014-2018)]

```
    ## iteration de la policy
    [~,~,~,~,Qon] = mdp_policy_iteration(transition, R, gamma, ones(ncl2,1))
        ##cf MDPToolbox Matlab
    [~,OptimalAction]=max(Qon,[],2) ##deterministic
    OA(:,modl)=OptimalAction ##save optimal actions

    ## evaluation off-policy

    ## add pi(s,a) and b(s,a)
    p=0.01 ## softening policies
    softpi=physpol ## behavior policy = clinicians'

    for (i in 1:750)
    {
        ii=softpi(i,:)==0
        z=p/sum(ii)
        nz=p/sum(~ii)
        softpi(i,ii)=z
        softpi(i,~ii)=softpi(i,~ii)-nz
    }
    softb=abs(zeros(752,25)-p/24); ## "optimal" policy = target policy =
       evaluation policy

    for( i in 1:750)
    {
        softb(i,OptimalAction(i))=1-p
    }


    for (i in 1:size(data))
    {
        data[i,nbOfFeatures+1]=OptimalAction(data[i,2])
    }
```

Table 1 | Description of the datasets

	MIMIC-III	eRI
Unique ICUs (*n*)	5	128
Unique ICU admissions (*n*)	17,083	79,073
Characteristics of hospitals, per number of ICU admissions	Teaching tertiary hospital	Nonteaching: 37,146 (47.0%) Teaching: 29,388 (37.2%) Unknown: 12,539 (15.9%)
Age, years (mean (s.d.))	64.4 (16.9)	65.0 (16.7)
Male gender (*n* (%))	9,604 (56.2%)	40,949 (51.8%)
Premorbid status (*n* (%))		
Hypertension	9,384 (54.9%)	43,365 (54.8%)
Diabetes	4,902 (28.7%)	25,290 (32.0%)
CHF	5,206 (30.5%)	15,023 (19.0%)
Cancer	1,803 (10.5%)	11,807 (14.9%)
COPD or RLD	4,248 (28.7%)	18,406 (23.3%)
CKD	3,087(18.1%)	14,553 (18.4%)
Primary ICD-9 diagnosis (*n* (%))		
Sepsis, including pneumonia	5,824 (34.1%)	41,396 (52.3%)
Cardiovascular	5,270 (30.8%)	11,221 (14.2%)
Respiratory	1,798 (10.5%)	9,127 (11.5%)
Neurological	1,590 (9.3%)	7,127 (9.0%)
Renal	429 (2.5%)	1,454 (1.8%)
Others	2,172 (12.7%)	8,747 (11.1%)
Initial OASIS (mean (s.d.))	33.5 (8.8)	34.8 (12.4)
Initial SOFA (mean (s.d.))	7.2 (3.2)	6.4 (3.5)
Procedures during the 72 h of data collection:		
Mechanical ventilation (*n* (%))	9,362 (54.8%)	39,115 (49.5%)
Vasopressors (*n* (%))	6,023 (35.3%)	23,877 (30.2%)
Renal replacement therapy (*n* (%))	1,488 (8.7%)	6,071 (7.7%)
Length of stay, days (median, (IQR))	3.1 (1.8-7)	2.9 (1.7-5.6)
ICU mortality	7.4%	9.8%
Hospital mortality	11.3%	16.4%
90-d mortality	18.9%	Not available

CHF, congestive heart failure; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; ICD-9, International Classification of Diseases version 9; IQR, interquartile range; OASIS, Oxford Acute Severity of Illness Score; RLD, restrictive lung disease; SOFA, sequential organ failure assessment.

Figure 5.1 – Principales variables descriptives des registres MIMIC-III et eRI

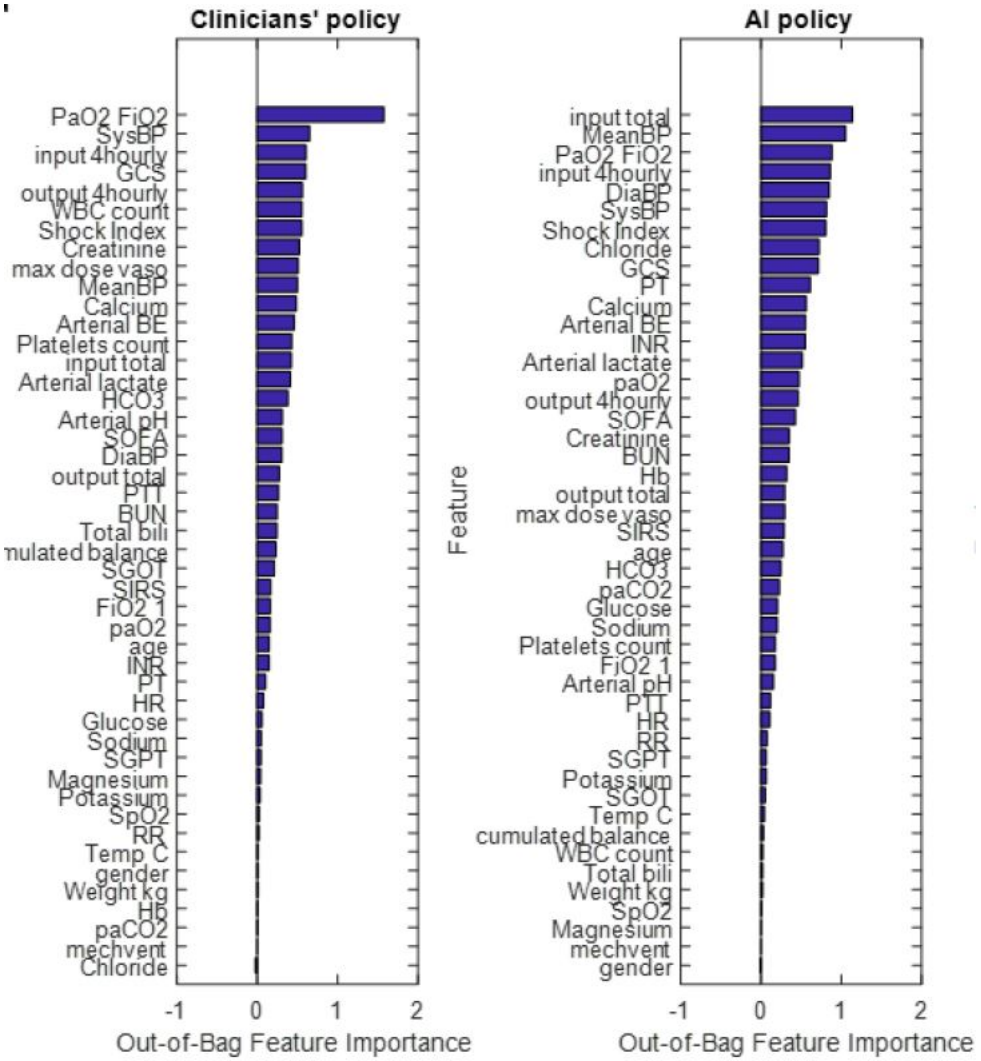

Figure 5.2 – Importance relative des paramètres dans la politique "médecins" et "AI"

Tableau 5.1 – Variables sélectionnées pour la construction du modèle

Variables	Type	Unité
'gender'	Binaire	-
'age'	Continue	jours
're_admission'	Binaire	-
'Weight_kg'	Continue	kg
'GCS'	Continue	-
'HR'	Continue	bpm
'SysBP'	Continue	mmHg
'MeanBP'	Continue	mmHg
'DiaBP'	Continue	mmHg
'RR'	Continue	bpm
'SpO2'	Continue	%
'Temp_C'	Continue	Celsius
'FiO2_1'	Continue	Fraction
'Potassium'	Continue	meq/L
'Sodium'	Continue	meq/L
'Chloride'	Continue	meq/L
'Glucose'	Continue	mg/dL
'BUN'	Continue	mg/dL
'Creatinine'	Continue	mg/dL
'Magnesium'	Continue	mg/dL
'Calcium'	Continue	mg/dL
'SGOT'	Continue	u/L
'SGPT'	Continue	u/L
'Total_bili'	Continue	mg/dL
'Hb'	Continue	g/dL
'WBC_count'	Continue	E9/L
'Platelets_count'	Continue	E9/L
'PTT'	Continue	s
'PT'	Continue	s
'INR'	Continue	-
'Arterial_pH'	Continue	-
'paO2'	Continue	mmHg
'paCO2'	Continue	mmHg
'Arterial_BE'	Continue	meq/L
'Arterial_lactate'	Continue	mmol/L
'HCO3'	Continue	meq/L
'mechvent'	Binaire	-
'Shock_Index'	Continue	bpm/mmHg
'PaO2_FiO2'	Continue	mmHg
'max_dose_vaso'	Continue	mcg/kg/min
'input_total_tev'	Continue	mL
'input_4hourly_tev'	Continue	mL
'output_total'	Continue	mL
'output_4hourly'	Continue	mL
'cumulated_balance_tev'	Continue	mL
'SOFA'	Integer	-
'SIRS'	Integer	-