Aprendizado por Reforço para Alocação de Rins: Uma Abordagem Baseada em *Policy Gradient* com Dados Reais do Sistema Brasileiro de Transplantes

Ivar V. Belizario[1], Douglas Teodoro[2], Luis G. M. Andrade[3], Gabriel Spadon[4], Jose F. Rodrigues Jr.[1]

[1] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil

[2]Department of Informatics, University of Geneva, Geneva, Switzerland

[3]Faculty of Medicine of Botucatu, Sao Paulo State University, Botucatu, Brazil

[4]Department of Computer Science, Dalhousie University, Halifax, Canada

ivar@alumni.usp.br, douglas.teodoro@unige.ch, gustavo.modelli@unesp.br, spadon@dal.ca, junio@icmc.usp.br

Resumo. *Este trabalho apresenta um agente de alocação de rins baseado em Aprendizado por Reforço Profundo (DRL), visando otimizar a sobrevida no sistema brasileiro de transplantes. O modelo foi treinado e avaliado em cenários pré e durante a COVID-19, demonstrando convergência estável. Os resultados indicam que a parametrização da recompensa permite ajustar a política conforme as prioridades clínicas: a configuração C1 reduziu a taxa de mortalidade em lista de 11,32% para 8,07% (sem pandemia) e de 10,04% para 6,67% (durante a pandemia), superando o sistema atual (BRAS) em todas as métricas e diminuindo o tempo médio de espera pandémico de 476,7 para 188,9 dias. Já a configuração C2 priorizou estritamente a mortalidade em lista, reduzindo-a para os mínimos de 7,59% (sem pandemia) e 6,36% (durante a pandemia), embora com um impacto severo de aumento no tempo de espera sob condições pandémicas.*

Abstract. *This work presents a kidney allocation agent based on Deep Reinforcement Learning (DRL) aimed at optimizing survival in the Brazilian transplant system. The model was trained and evaluated under pre- and during COVID-19 scenarios, showing stable convergence. The results indicate that reward parameterization allows for flexible policy adjustments according to clinical priorities: configuration C1 reduced waiting-list mortality from 11.32% to 8.07% (pre-pandemic) and from 10.04% to 6.67% (during the pandemic), outperforming the current system (BRAS) across all metrics and cutting the average pandemic waiting time from 476.7 to 188.9 days. Meanwhile, configuration C2 strictly prioritized list mortality, minimizing it to 7.59% (pre-pandemic) and 6.36% (during the pandemic), albeit at the cost of a severe increase in waiting times under pandemic conditions.*

1. Introdução

O transplante renal é o tratamento mais eficaz para pacientes com doença renal crônica em estágio terminal, proporcionando melhor qualidade de vida e maior sobrevida em

comparação à diálise [Matas et al. 2023]. Entretanto, a escassez de órgãos e o crescimento contínuo das listas de espera permanecem entre os principais desafios dos sistemas de saúde [Kupiec-Weglinski 2022]. Como consequência, muitos pacientes aguardam longos períodos por um transplante e parte deles falece antes da disponibilidade de um órgão [Tonelli et al. 2011].

No Brasil, o *Sistema Nacional de Transplantes* (SNT) coordena a captação, distribuição e monitoramento dos órgãos doados. Embora o país possua um dos maiores programas públicos de transplantes do mundo, fatores como desigualdades regionais, limitações de infraestrutura e a complexidade dos critérios de priorização dificultam o processo de alocação [Okano et al. 2023]. Atualmente, a distribuição de rins considera grupo sanguíneo, compatibilidade HLA (Antígenos Leucocitários Humanos), tempo em lista e critérios de urgência médica, exigindo um equilíbrio entre eficiência e equidade [Elalouf and Pliskin 2022, Lima et al. 2023].

Nesse contexto, técnicas de *inteligência artificial* (IA) vêm sendo investigadas como ferramentas de apoio à decisão [Li et al. 2023, Ding et al. 2025, Naqvi et al. 2025]. Modelos supervisionados têm sido utilizados para prever desfechos pós-transplante e compatibilidade entre doador e receptor [Deshpande 2024], enquanto abordagens de otimização multiobjetivo e sistemas de apoio à decisão buscam integrar critérios clínicos e éticos [Jalilvand et al. 2023, Dale et al. 2024]. Entretanto, essas abordagens tendem a operar de forma estática, com capacidade limitada para se adaptar às mudanças na composição da lista de espera e na disponibilidade de órgãos.

Métodos de *aprendizado por reforço* (*Reinforcement Learning* – RL) oferecem uma alternativa para esse cenário por modelarem o problema como um processo sequencial de decisão [Sutton and Barto 2018, Deshpande 2024]. Nesse paradigma, um agente aprende uma política de alocação por meio da maximização de recompensas associadas a critérios de interesse. Estudos recentes mostram que algoritmos como *Deep Q-Learning*, *Actor-Critic* e *Policy Gradient* são capazes de capturar aspectos temporais e probabilísticos presentes na tomada de decisão médica [Naqvi et al. 2025, Ding et al. 2025].

Este trabalho investiga a seguinte questão de pesquisa: *um agente de aprendizado por reforço treinado com dados reais do Sistema Nacional de Transplantes brasileiro é capaz de aprender uma política de alocação renal que reduza a mortalidade em lista de espera e o tempo médio de espera, preservando a probabilidade de sobrevida pós-transplante?*

Neste estudo, o conceito de equidade é definido como o equilíbrio entre maximização da sobrevida pós-transplante e priorização de pacientes com maior risco de morte e maior tempo de espera. Essa interpretação está alinhada a propostas recentes que defendem a incorporação explícita de critérios de risco em políticas de alocação de órgãos [Dale et al. 2024, Zhu et al. 2024].

Para responder à questão proposta, desenvolveu-se uma abordagem baseada em *Policy Gradient* aplicada ao contexto do SNT. Diferentemente de estudos baseados em dados simulados ou registros internacionais, o método utiliza dados reais do sistema brasileiro, abrangendo períodos anteriores e posteriores à pandemia de COVID-19.

As principais contribuições deste trabalho são:

- aplicação de uma abordagem baseada em *Policy Gradient* ao problema de alocação renal utilizando dados reais do Sistema Nacional de Transplantes brasileiro;
- formulação de uma função de recompensa que combina probabilidade de sobrevida pós-transplante, probabilidade de morte em lista de espera e tempo de espera;
- construção de um ambiente de treinamento baseado em uma lista de espera dinâmica reconstruída temporalmente;
- avaliação da política aprendida em cenários pré-pandemia e durante a pandemia de COVID-19.

Os resultados mostram que o aprendizado por reforço pode atuar como mecanismo de apoio à decisão em transplantes, contribuindo para políticas de alocação mais eficientes e equitativas. As seções seguintes apresentam a metodologia proposta, os dados utilizados e os resultados experimentais.

2. Trabalhos relacionados

O *Deep Reinforcement Learning* (DRL) tem sido aplicado a problemas de jogos, visão computacional, navegação e otimização [Sá and Madeira 2025, Zhu et al. 2025, Zhao et al. 2024, Tang et al. 2025]. De forma geral, os métodos podem ser divididos em duas categorias: métodos baseados em funções de valor, como o *Deep Q-Learning*, e métodos baseados em política, como o *Policy Gradient*. Na alocação de órgãos, abordagens baseadas em política são particularmente adequadas por aprenderem diretamente a política de decisão a partir de uma função de recompensa.

Naqvi et al. [Naqvi et al. 2025] propõem uma abordagem de aprendizado por reforço para alocação de rins que considera simultaneamente sobrevivência pós-transplante e critérios éticos. O problema é formulado como um processo de decisão de Markov (MDP), no qual o estado representa a lista de espera e o doador disponível, enquanto a ação corresponde à escolha do receptor. A função de recompensa combina benefício líquido, espera prolongada e mortalidade em lista. Utilizando aproximadamente 5.000 pares doador–receptor derivados do Scientific Registry of Transplant Recipients (SRTR, EUA), os autores demonstram reduções na mortalidade em lista e no tempo médio de espera, além de ganhos em sobrevivência pós-transplante e benefício líquido quando comparados a políticas tradicionais, como *first-come, first-served* (FCFS, primeiro a chegar, primeiro a ser atendido), *utility-first* e *benefit-first* [Naqvi et al. 2025].

Ding et al. [Ding et al. 2025] apresentam o *FairAlloc*, uma abordagem para alocação de fígados baseada em *Policy Gradient*. O método modela a alocação como um problema de ranqueamento multiobjetivo que busca maximizar a sobrevivência do enxerto e minimizar desigualdades entre pacientes. Avaliado com dados do Organ Procurement and Transplantation Network (OPTN) [Matas et al. 2023], o modelo foi comparado a seis métodos de referência e demonstrou que critérios explícitos de equidade podem ser incorporados diretamente ao processo de decisão [Ding et al. 2025].

Além do aprendizado por reforço, a literatura inclui modelos supervisionados, métodos de otimização e sistemas de apoio à decisão. Modelos supervisionados são amplamente utilizados para estimar risco, compatibilidade e sobrevida pós-transplante. Métodos de otimização multiobjetivo permitem representar simultaneamente critérios clínicos e éticos, enquanto sistemas de apoio à decisão oferecem maior interpretabilidade

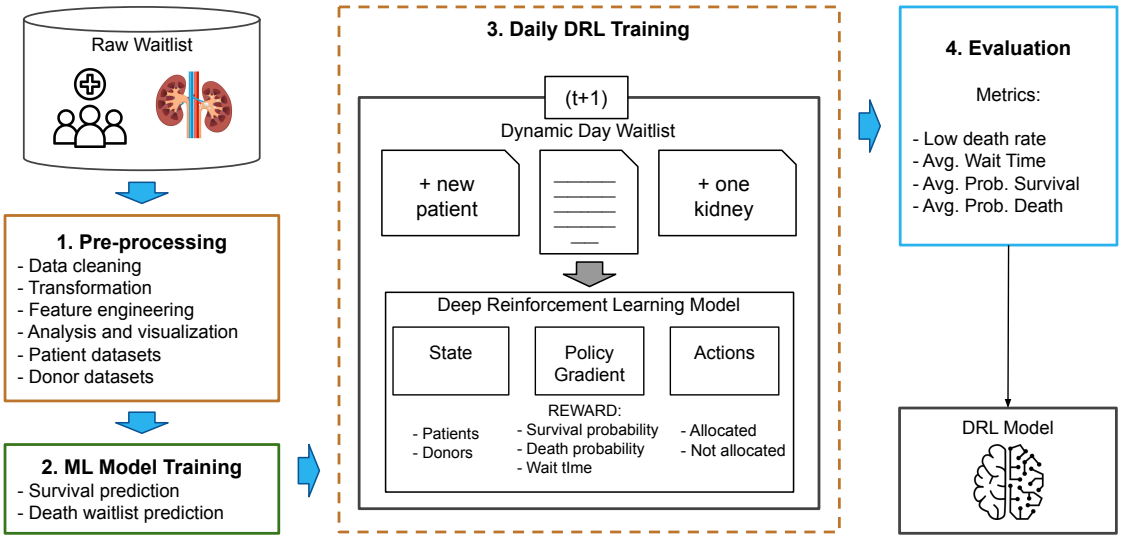

Figura 1. Metodologia proposta para o processo de alocação de rins baseado no treinamento de modelos de aprendizado por reforço profundo.

operacional. Entretanto, essas abordagens normalmente não modelam explicitamente a natureza sequencial do processo de alocação nem atualizam decisões de forma adaptativa à medida que a lista de espera evolui.

Portanto, a contribuição deste trabalho não está na proposição de um novo algoritmo de aprendizado por reforço. Métodos baseados em *Policy Gradient* já foram explorados em transplantes de órgãos [Naqvi et al. 2025, Ding et al. 2025]. A contribuição reside na adaptação dessa classe de métodos ao contexto do Sistema Nacional de Transplantes brasileiro, utilizando dados reais, uma lista de espera reconstruída dinamicamente e uma função de recompensa que combina probabilidade de sobrevida pós-transplante, probabilidade de morte em lista de espera e tempo de espera. Essa formulação permite avaliar políticas adaptativas em um contexto operacional distinto dos cenários previamente estudados.

3. Método proposto

A metodologia proposta é formulada como uma solução computacional baseada em inteligência artificial, considerando os princípios de alocação com equidade discutidos nos trabalhos relacionados sobre transplante de órgãos.

As etapas da metodologia são ilustradas na Figura 1. Na Etapa 1, a partir da lista de espera, são executadas tarefas de pré-processamento das características dos pacientes. Na Etapa 2, são treinados modelos clássicos de aprendizado de máquina (ML) para estimar a probabilidade de sobrevivência após o transplante e a probabilidade de morte na lista de espera. Em seguida, na Etapa 3, a lista de espera é dividida dinamicamente por dia de alocação, a fim de treinar um modelo baseado em aprendizado por reforço profundo (DRL). Por fim, na Etapa 4, o modelo é avaliado por meio de métricas específicas para analisar seu desempenho e capacidade de aprendizado.

3.1. Conjunto de dados - Lista de espera

O conjunto de dados da lista de espera original [Salomão Pontes et al. 2024] foi disponibilizado pela equipe do Hospital das Clínicas da Faculdade de Medicina de Botucatu da

Universidade Estadual Paulista (HCFMB-UNESP).

O conjunto de dados da lista de espera, disponibilizado no formato XLSX, é composto por aproximadamente 154.436 registros de pacientes, registrados entre 1 de janeiro de 2012 e 31 de dezembro de 2022, incluindo pacientes registrados durante o período da pandemia. Entre os atributos, estão dados demográficos do paciente (idade no momento do registro, sexo, cor/raça e unidade federativa de residência); dados clínicos (tipo sanguíneo, número de transfusões, painel de reatividade e doença base); dados de localização do centro de saúde responsável pelo atendimento (unidade federativa e laboratório); atributos temporais (data de nascimento, data de registro na lista de espera, data de alocação, data de óbito e indicador de registro durante a pandemia); e atributos-alvo, que indicam se o paciente recebeu a alocação do órgão ou se veio a óbito enquanto aguardava na lista de espera.

Além disso, como a base de dados da lista de espera não contempla diretamente informações dos doadores, os registros de pacientes que receberam alocação foram utilizados para inferir as características dos doadores e para reconstruir a lista de espera de forma dinâmica, organizada por dia de alocação, no qual podem ocorrer múltiplas alocações quando há mais de um doador no dia.

Cabe destacar que o treinamento do modelo de aprendizado por reforço é realizado considerando **estados** compostos pelas informações e características dos pacientes e dos doadores. As **ações** possíveis correspondem à alocação ou à não alocação do rim, enquanto os **episódios** de treinamento compreendem os dias de alocação, nos quais uma lista dinâmica de espera é avaliada para determinar qual paciente receberá o órgão.

3.2. Pré-processamento

As tarefas de pré-processamento incluíram limpeza, transformação, engenharia de características (*feature engineering*), análise e visualização, além da criação de dois conjuntos de dados. Na etapa de limpeza, foram eliminadas instâncias incompletas, duplicadas ou com dados inconsistentes. Na transformação, dados categóricos foram convertidos para representações numéricas ou binárias. A codificação *one-hot encoding* foi aplicada às variáveis categóricas sem ordenação natural, como sexo, raça/cor, doença base, unidade federativa, laboratório e tipo sanguíneo, quando utilizadas como entrada dos modelos.

Na etapa de engenharia de características, foram criadas novas características a partir de variáveis já existentes. Incluiu-se uma característica binária para diálise e diabetes, bem como três características categóricas compostas: raça e doença, sexo e raça, e sexo e doença. A análise e visualização dos dados foram utilizadas para distinguir pacientes que sobreviveram ao transplante daqueles que faleceram na lista de espera, bem como para compreender a distribuição das variáveis e identificar possíveis desbalanceamentos.

Após o pré-processamento, foram definidos dois conjuntos de dados de pacientes: o primeiro conjunto, denominado DS1, não inclui pacientes nem alocações durante a pandemia; o segundo, denominado DS2, inclui pacientes e alocações durante a pandemia. A Tabela 1 resume os dois conjuntos de dados gerados. Como a lista de espera original não contemplava diretamente os doadores nem suas respectivas características, foram considerados os pacientes que receberam uma alocação para inferir dados dos doadores. Nesse processo, foi atribuído aos doadores o tipo sanguíneo correspondente ao dos pacientes. Portanto, nas tarefas de alocação, a compatibilidade é avaliada pelo grupo sanguíneo.

Tabela 1. Definição dos dois conjuntos de dados utilizados: DS1, que exclui pacientes e doadores durante a pandemia; e DS2, que inclui pacientes e doadores durante a pandemia.

Conjunto	Início	Fim	Pacientes	Doadores
DS1	2012-01-01	2019-12-31	84.680	26.901
DS2	2012-01-01	2022-12-31	118.498	39.428

3.3. Treinamento de modelos de ML

Nesta etapa, foram treinados modelos de classificação supervisionada para estimar a probabilidade de sobrevivência pós-transplante e a probabilidade de morte na lista de espera. Essas estimativas são posteriormente incorporadas à função de recompensa utilizada pelo modelo de aprendizado por reforço.

O problema foi formulado como uma tarefa de classificação binária, na qual cada instância representa o estado clínico de um paciente em um determinado momento da lista de espera. As variáveis de entrada correspondem a atributos demográficos, clínicos e temporais, enquanto a variável alvo representa o desfecho observado ao final do acompanhamento. Após o treinamento, os modelos passaram a fornecer probabilidades associadas aos dois desfechos de interesse.

Os experimentos foram conduzidos sobre os conjuntos DS1 e DS2, utilizando divisão de 70% para treinamento e 30% para teste. Como as classes apresentavam desbalanceamento, foram aplicadas técnicas de sobre-amostragem apenas no conjunto de treinamento. Foram avaliados os métodos SMOTE, ADASYN, KMeansSMOTE, BorderlineSMOTE e RandomOverSampler. Cada técnica gerou um conjunto balanceado distinto, utilizado no treinamento e posteriormente avaliado sobre o conjunto de teste original.

Foram avaliados três algoritmos de classificação: Random Forest (RF), Gradient Boosting Classifier (GBC) e XGBoost Classifier (XGBC). O RF foi configurado com 550 árvores, critério de Gini e profundidade máxima de 35. O GBC utilizou 100 estimadores, *learning rate* de 0,1 e profundidade máxima de 3. O XGBC foi treinado com 250 árvores, *learning rate* de 0,33, profundidade máxima de 3 e regularização padrão.

O desempenho foi avaliado por meio de F1-score, precisão e revocação (*recall*), calculadas no conjunto de teste. Em ambos os conjuntos de dados, o XGBoost apresentou os melhores resultados e foi selecionado para estimar as probabilidades de sobrevivência pós-transplante e de morte na lista de espera. Essas probabilidades foram então utilizadas como variáveis de entrada no treinamento do modelo de aprendizado por reforço.

3.4. Treinamento do modelo de aprendizado por reforço

Nesta etapa, para cada conjunto DS1 e DS2, são utilizadas as informações dos doadores, ordenadas de forma crescente pela data de transplante. O conjunto de dados pré-processado é então dividido dinamicamente, filtrando, a cada dia, os pacientes cuja data de inscrição seja igual ou anterior à data registrada de um dado transplante. Durante essa execução dinâmica, o algoritmo de aprendizado por reforço é treinado continuamente, incluindo a atualização das características dos pacientes, tais como idade, tempo de espera e probabilidades de sobrevivência após o transplante e de morte na lista de espera.

De acordo com a revisão da literatura, os modelos baseados em *Policy Gradient* [Naqvi et al. 2025, Ding et al. 2025] demonstram melhor adequação ao contexto de transplantes de órgãos para pacientes em lista de espera. Dessa forma, foi implementado um modelo de aprendizado profundo por reforço baseado na estratégia *Policy Gradient*.

Na implementação proposta, foi desenvolvida uma política de alocação que considera as probabilidades de sobrevivência após o transplante, a probabilidade de morte na lista de espera e o tempo de permanência na lista. Essa política é formalizada na função de recompensa apresentada na Eq. 1.

$$R(i) = \alpha(ps(i)) + \theta(pd(i)) + \beta(wt(i)) \tag{1}$$

onde a função $ps(i)$ retorna a *probabilidade de sobrevivência após o transplante* do paciente i, enquanto $pd(i)$ representa a *probabilidade de morte do paciente enquanto aguarda a alocação do rim*. A função $wt(i)$ indica o *tempo total de espera* do paciente na lista. Os parâmetros α, θ e β são *coeficientes de ponderação*, com valores no intervalo $[0, 1]$, os quais ajustam a *relevância relativa de cada característica* durante o processo de maximização da função de recompensa. Essa formulação busca equilibrar critérios de maximização da sobrevida pós-transplante e priorização de pacientes com maior risco de morte e maior tempo de espera.

A formulação da função de recompensa R foi projetada com o objetivo de maximizar simultaneamente a seleção de pacientes com alta probabilidade de sobrevivência após o transplante e daqueles com elevada probabilidade de morte na lista de espera aguardando a alocação.

É importante destacar que as características associadas às probabilidades de sobrevivência e de morte, bem como a idade dos pacientes, são atualizadas dinamicamente a cada dia de avaliação para alocação, simulando uma fila de espera real ao longo do tempo. Após essa atualização, a função de recompensa é computada durante o processo de treinamento do modelo, permitindo que o modelo aprenda a priorizar as decisões com base em informações atualizadas.

Além disso, o critério de compatibilidade adotado considera os quatro tipos sanguíneos do sistema ABO — A, B, AB e O —, de modo que a alocação somente ocorre entre pacientes e doadores compatíveis.

3.5. Avaliação do modelo

Após a fase de treinamento, o algoritmo de aprendizado por reforço profundo é executado utilizando os pesos previamente ajustados, ou seja, sem atualização dos parâmetros.

Nesta etapa, são calculadas quatro métricas de avaliação com o objetivo de mensurar o desempenho e a capacidade de generalização do modelo: taxa de mortalidade, tempo médio de espera, probabilidade média de sobrevivência e probabilidade média de morte em lista. As métricas são definidas nas Eqs. 2–5, com base nos dados dos pacientes selecionados para a alocação de rim em cada momento no qual um órgão estava disponível segundo a sequência temporal observada na lista de espera.

$$\text{Taxa de Mortalidade} = \frac{1}{N} \sum_{i} \sigma(i) \tag{2}$$

$$\text{Tempo Médio de Espera} = \frac{1}{n}\sum_{i} wt(i) \tag{3}$$

$$\text{Prob. Média de Sobrevivência} = \frac{1}{n}\sum_{i} ps(i) \tag{4}$$

$$\text{Prob. Média de Morte} = \frac{1}{n}\sum_{i} pd(i) \tag{5}$$

onde a Eq. 2 representa a taxa de mortalidade, em que N é o número total de pacientes na lista de espera, e $\sigma(\cdot)$ é uma função indicadora que verifica se um paciente faleceu durante o período de espera. O parâmetro n corresponde ao número total de pacientes que receberam a alocação de um rim. A Eq. 3 define o tempo médio de espera, sendo $wt(\cdot)$ uma função que retorna o tempo de espera atual de cada paciente na lista. A Eq. 4 descreve a probabilidade média de sobrevivência, onde $ps(\cdot)$ fornece a probabilidade de sobrevivência estimada de cada paciente. Por fim, a Eq. 5 apresenta a probabilidade média de morte na lista de espera, em que $pd(\cdot)$ representa a probabilidade de morte associada a um determinado paciente.

3.6. Implementação

Os algoritmos de pré-processamento e a construção do modelo de aprendizado profundo, incluindo a função de recompensa que representa a política de alocação de rins, foram implementados na linguagem Python. As principais bibliotecas utilizadas foram: pandas (`https://pandas.pydata.org/`), para pré-processamento de dados; scikit-learn (`https://scikit-learn.org/`), para a implementação de modelos de aprendizado de máquina tradicional; e TensorFlow (`https://www.tensorflow.org/`), para a implementação do modelo de aprendizado por reforço profundo. O código-fonte será disponibilizado no repositório (`https://ivarvb.github.io/ka/`).

Todos os experimentos foram realizados em um servidor com aceleração por GPU. O treinamento foi executado em uma GPU NVIDIA Tesla P100, equipada com 16 GB de memória HBM2 e baseada na arquitetura Pascal. Essa GPU oferece alta largura de banda de memória e bom desempenho em operações de ponto flutuante de precisão simples (FP32), sendo adequada para o treinamento de modelos de aprendizado profundo.

4. Resultados experimentais e discussão

Após o pré-processamento dos dados, foram realizados experimentos relacionados à análise dos parâmetros da função de recompensa (Eq. 1) durante as fases de treinamento do modelo de aprendizado por reforço profundo, bem como à avaliação de sua eficiência na fase de teste. Esses experimentos foram conduzidos sobre dois conjuntos de dados, contendo pacientes e doadores em contextos sem considerar e considerando a pandemia, denominados DS1 e DS2, respectivamente. Para avaliar o processo de aprendizagem do modelo, foram executadas duas configurações experimentais, detalhadas a seguir.

4.1. Experimentos

A primeira configuração (C1) consiste em atribuir valores elevados aos parâmetros α e θ, especificamente 0.34, com o objetivo de conferir maior relevância às variáveis associadas

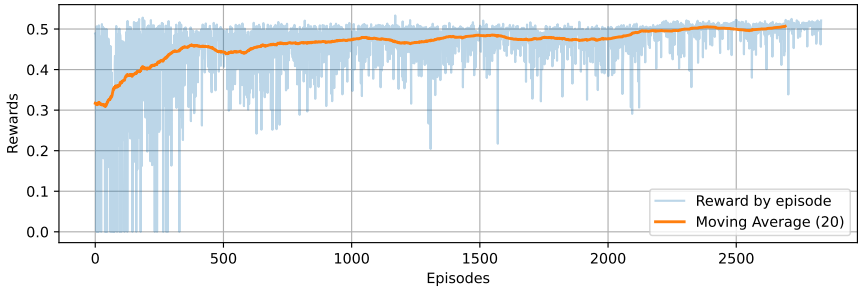

Figura 2. Resultados do treinamento da maximização da função de recompensa proposta. O gráfico corresponde à configuração C1 para o conjunto DS1.

à probabilidade de sobrevivência após o transplante e ao atendimento de pacientes com maior probabilidade de morte na lista de espera. Já, o parâmetro β é configurado com o valor 0.32, menor que os demais, reduzindo a importância do tempo de espera na alocação de rins. O objetivo deste experimento é avaliar se o modelo consegue aprender a maior importância dos parâmetros relacionados às probabilidades e a menor relevância do tempo de espera, favorecendo uma alocação de rins para pacientes com alta probabilidade de sobrevivência e elevado risco de morte na lista de espera.

Na segunda configuração (C2), os parâmetros são configurados da seguinte forma: os parâmetros α e θ também são definidos com valores elevados, iguais a 0.32, a fim de enfatizar a relevância das probabilidades de sobrevivência após o transplante e do atendimento a pacientes com maior probabilidade de morte na lista de espera. Especificamente, o parâmetro β é configurado com o valor 0.36, superior aos demais parâmetros, com o objetivo de atribuir maior importância ao tempo de espera na alocação de rins. O objetivo deste experimento é avaliar se o modelo é capaz de aprender a maior importância do tempo de espera e verificar como essa configuração influencia as demais métricas de avaliação.

4.2. Resultados

A Figura 2 ilustra o comportamento do modelo durante o processo de treinamento, mostrando a evolução das recompensas acumuladas ao longo dos episódios. Nota-se que, tanto para os conjuntos sem pandemia quanto com pandemia, o modelo apresenta convergência estável e crescimento progressivo das recompensas, indicando aprendizado consistente da política de alocação. Essa estabilidade reflete a capacidade do modelo em identificar padrões relevantes de priorização entre pacientes e doadores, mesmo em cenários de alta variabilidade. A convergência suave observada em ambas as configurações (C1 e C2) também sugere robustez do método e ausência de sobreajuste, demonstrando potencial de generalização da política aprendida para diferentes contextos temporais e epidemiológicos. Conclui-se que houve convergência do modelo no sentido de aprender as melhores ações para a alocação.

A Tabela 2 apresenta os resultados obtidos na fase de teste, comparando as políticas aprendidas com os dados reais do sistema brasileiro de alocação (BRAS). Verifica-se que ambas as configurações propostas reduziram de forma significativa as taxas de mortalidade em lista. No cenário sem pandemia, a taxa caiu de 0.1132 (BRAS) para 0.0807 (C1) e 0.0759 (C2); no cenário com pandemia, a redução foi de 0.1004 para 0.0667 e 0.0636, respectivamente.

Tabela 2. Resultados do experimento de alocação de rins utilizando quatro métricas de avaliação. A configuração de parâmetros da função de recompensa segue duas configurações: C1 ($\alpha = 0.34$, $\theta = 0.34$ e $\beta = 0.32$) e C2 ($\alpha = 0.32$, $\theta = 0.32$ e $\beta = 0.36$).

Métrica	Sem pandemia			Com pandemia			Ótimo
	BRAS	$\mathbf{R}$ (C1)	$\mathbf{R}$ (C2)	BRAS	$\mathbf{R}$ (C1)	$\mathbf{R}$ (C2)	
Taxa de mortalidade	0.1132	0.0807	**0.0759**	0.1004	0.0667	**0.0636**	$\Downarrow$
Tempo médio de espera	410.6911	**280.3532**	300.9115	476.6967	**188.8750**	954.3657	$\Downarrow$
Prob. média de sobrevivência	0.7902	**0.8821**	0.8631	0.8139	**0.9155**	0.8925	$\Uparrow$
Prob. média de morte	0.8611	**0.9670**	0.9649	0.7119	**0.9337**	0.7584	$\Uparrow$

4.3. Discussão

Os resultados indicam que o modelo aprendeu uma política de alocação capaz de reduzir simultaneamente a taxa de mortalidade e o tempo médio de espera em comparação ao sistema brasileiro de alocação (BRAS). A redução observada nessas duas métricas sugere que a política aprendida passou a priorizar pacientes com maior risco de morte em lista sem comprometer a eficiência do processo de alocação.

A configuração C1 reduziu o tempo médio de espera de 410,69 para 280,35 dias no cenário sem pandemia e de 476,70 para 188,88 dias no cenário com pandemia. Esses resultados indicam que a política aprendida pode acelerar o acesso ao transplante, reduzindo a permanência dos pacientes em diálise e os riscos associados à espera prolongada.

A redução da mortalidade também não ocorreu à custa de uma diminuição da probabilidade média de sobrevida pós-transplante. Em ambas as configurações, essa métrica apresentou valores superiores aos observados no BRAS. Isso sugere que o modelo foi capaz de combinar critérios de urgência e benefício clínico, evitando priorizações baseadas exclusivamente em um único objetivo.

A comparação entre as configurações evidencia o efeito dos pesos da função de recompensa. Em C2, o aumento do peso associado ao tempo de espera ($\beta = 0.36$) levou a mudanças substanciais no comportamento da política, especialmente no cenário pandêmico, em que o tempo médio de espera atingiu 954,36 dias. Esse resultado mostra que a definição dos pesos influencia diretamente os compromissos entre mortalidade, tempo de espera e benefício clínico, devendo ser tratada como uma decisão de política pública e não apenas como um ajuste técnico.

De forma geral, C1 apresentou o comportamento mais equilibrado entre as métricas avaliadas, obtendo simultaneamente baixa mortalidade, menor tempo de espera e maiores probabilidades médias de sobrevivência. Já C2 alcançou as menores taxas de mortalidade, porém com perda de desempenho em outras métricas, particularmente no cenário com pandemia. Esses resultados mostram que a formulação proposta é flexível para representar diferentes prioridades de alocação por meio da configuração da função de recompensa.

Por fim, a estabilidade observada durante o treinamento e os ganhos obtidos em relação ao BRAS sugerem que a abordagem possui potencial para apoiar a formulação de políticas de alocação mais eficientes. Entretanto, sua aplicação prática exigiria calibração dos parâmetros da recompensa, validação prospectiva e análises adicionais de equidade em diferentes subgrupos populacionais.

5. Conclusões

Este trabalho apresentou uma metodologia baseada em aprendizado por reforço, especificamente *Policy Gradient*, para apoiar a alocação de rins com dados reais do Sistema Nacional de Transplantes brasileiro. Os resultados indicam que o modelo aprende uma política de alocação compatível com a função de recompensa proposta, que combina probabilidade de sobrevivência após o transplante, probabilidade de morte em lista de espera e tempo de espera. Em comparação com os dados reais de alocação do BRAS, a política aprendida reduziu a taxa de mortalidade em lista e o tempo médio de espera, ao mesmo tempo em que priorizou pacientes com maior probabilidade de sobrevivência pós-transplante e maior risco de morte em lista.

A análise comparativa entre C1, que prioriza probabilidade de sobrevida pós-transplante e risco de morte em lista, e C2, que atribui maior peso ao tempo de espera, mostrou que os pesos da função de recompensa influenciam diretamente o comportamento da política aprendida. C1 apresentou comportamento mais equilibrado entre mortalidade, tempo de espera e probabilidade estimada de sobrevivência, enquanto C2 obteve as menores taxas de mortalidade. Esse resultado indica que a formulação proposta permite ajustar a política de alocação conforme diferentes prioridades clínicas e operacionais. A consistência dos resultados nos cenários pré-pandemia e durante a pandemia sugere que a abordagem é robusta a mudanças no contexto de alocação. Ainda assim, sua adoção prática exigiria validação adicional, calibração dos pesos da recompensa e análise de equidade por subgrupos, como região, raça/cor, comorbidades e vulnerabilidade social.

Como trabalhos futuros, serão investigadas novas formulações da função de recompensa para representar diferentes políticas de alocação, incluindo critérios de vulnerabilidade, comorbidades e desigualdades regionais. Também serão avaliados modelos mais robustos para estimar as probabilidades de sobrevivência após o transplante e de mortalidade em lista de espera.

Agradecimentos

Os autores agradecem o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – processos 2024/04761-0, 2019/07665-4 e 2024/13328-9), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – processo 304805/2025-4), da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Código de Financiamento 001) e do Swiss National Science Foundation (SNSF).

Referências

[Dale et al. 2024] Dale, R., Cheng, M., Casselman Pines, K., and Currie, M. (2024). Inconsistent values and algorithmic fairness: a review of organ allocation priority systems in the united states. *BMC Medical Ethics*, 25.

[Deshpande 2024] Deshpande, R. (2024). Smart match: revolutionizing organ allocation through artificial intelligence. *Frontiers in Artificial Intelligence*, Volume 7 - 2024.

[Ding et al. 2025] Ding, S., Zha, D., Zhang, K., Chen, L., Jiang, X., Hu, X., and Zou, N. (2025). Fairalloc: Learning fair organ allocation policy for liver transplant. *Journal of Healthcare Informatics Research*.

[Elalouf and Pliskin 2022] Elalouf, A. and Pliskin, J. S. (2022). Balancing equity and efficiency in kidney allocation: An overview. *Cambridge Quarterly of Healthcare Ethics*, 31(3):321–332.

[Jalilvand et al. 2023] Jalilvand, N., Bairamzadeh, S., Tavakkoli-Moghaddam, R., and Azaron, A. (2023). A bi-objective organ transplant supply chain network with recipient priority considering carbon emission under uncertainty. *Computers & Industrial Engineering*, 181:109295.

[Kupiec-Weglinski 2022] Kupiec-Weglinski, J. W. (2022). Grand challenges in organ transplantation. *Frontiers in Transplantation*, 1:897679.

[Li et al. 2023] Li, H., Zhang, W., and Chen, L. (2023). Predicting kidney transplant outcomes using machine learning: A systematic review. *Frontiers in Medicine*, 10:1158704.

[Lima et al. 2023] Lima, B. A., Reis, F., Alves, H., and Henriques, T. S. (2023). Equity matrix for kidney transplant allocation. *Transplant Immunology*, 81:101917.

[Matas et al. 2023] Matas, A. J., Smith, D. L., Skeans, J. A., and Stewart, J. P. (2023). Optn/srtr 2022 annual data report: Kidney. *American Journal of Transplantation*, 23(S1):1–49.

[Naqvi et al. 2025] Naqvi, S. A. A., Tennankore, K., Worthen, G., Vinson, A., and Abidi, S. S. R. (2025). A reinforcement learning framework for optimizing kidney allocation for transplant based on survival and ethical criteria. In *Artificial Intelligence in Medicine*, pages 323–332. Springer Nature.

[Okano et al. 2023] Okano, C. d. S., Menezes, C. C. S. d., Brandão, F. A., Carletto, V. R., and de Souza, M. A. (2023). Analysis of the national transplant scenario in brazil. *Research, Society and Development*, 12(9).

[Sá and Madeira 2025] Sá, G. C. B. e. and Madeira, C. A. G. (2025). Deep reinforcement learning in real-time strategy games: a systematic literature review. *Applied Intelligence*, 55(3):243.

[Salomão Pontes et al. 2024] Salomão Pontes, D. F., Fernandes Ferreira, G., Segev, D., Massie, A. B., Levan, M., Barbosa, A. M. P., da Rocha, N. C., and Modelli de Andrade, L. G. (2024). Regional disparities in kidney transplant allocation in brazil: A retrospective cohort study. *Clinical Transplantation*, 38(9):e15446.

[Sutton and Barto 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.

[Tang et al. 2025] Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., and Stone, P. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8:153–188.

[Tonelli et al. 2011] Tonelli, M., Wiebe, N., Knoll, G., and et al. (2011). Kidney transplantation compared with dialysis in clinically relevant outcomes: a systematic review. *The Lancet*, 378(9800):180–189.

[Zhao et al. 2024] Zhao, D., Huanshi, X., and Zhang, X. (2024). Active exploration deep reinforcement learning for continuous action space with forward prediction. *International Journal of Computational Intelligence Systems*, 17.

[Zhu et al. 2024] Zhu, L., Gao, W., Zhang, S., et al. (2024). Equality of healthcare resource allocation between impoverished counties and non-impoverished counties in northwest china: a longitudinal study. *BMC Health Services Research*, 24.

[Zhu et al. 2025] Zhu, Y. et al. (2025). Deep reinforcement learning of mobile robot navigation: A comparative analysis of value-based, policy-based and hybrid methods. *Sensors*, 25(11).