Offline-RL 기반 패혈증 치료 연구의 한계와 재현성·안전성·일반화를 위한 체크리스트

성다훈[1], 채송화[2], 임유진[3]
[1]숙명여자대학교 IT공학과 박사과정
[2]숙명여자대학교 IT공학과 석사과정
[3]숙명여자대학교 인공지능공학부 교수
{ekgns324, watermelon97, yujin91}@sookmyung.ac.kr

Limitations of Offline-RL-Based Sepsis Treatment Research and Checklist for Reproducibility, Safety, and Generalization

Da-Hun Seong[1], Song-Hwa Chae[2], Yujin Lim[3]
[1,2]Dept. of Information Technology Engineering, Sookmyung Women's University
[3]Div. of Artificial Intelligence Engineering, Sookmyung Women's University

요 약

패혈증은 전 세계적으로 높은 사망률을 보이는 중증 질환으로, 패혈증 환자 치료 전략의 최적화를 위해 Offline-RL이 연구에 활용되어 왔다. 그러나 기존의 Offline-RL 기반 패혈증 치료 연구의 주요 한계가 체계적으로 분석되지 않아, 임상 적용 시 안전성과 일반화를 위한 통합된 연구에 어려움이 존재한다. 따라서 본 연구는 이러한 한계를 분석하고, 연구의 재현성·안전성·일반화 관점에서 고려해야 할 체크리스트를 제안한다. 이를 통해 연구자와 임상의가 RL 모델을 개발 및 평가할 때 최소한의 안전성 기준을 점검할 수 있도록 돕는다.

1. 서론

패혈증은 감염에 대한 숙주의 조절 장애로 발생하는 생명을 위협하는 장기 기능 부전 상태로 정의되며[1], 전 세계적으로 높은 사망률을 보이는 중증 질환이다. 조기 치료가 예후를 좌우하나, 임상 경과가 복잡하고 환자별 이질성이 커, 최적 치료 전략에 어려움이 있다. 특히 중환자실(Intensive Care Unit, ICU) 환경에서는 연속적인 의사결정이 필수적이다. 이러한 맥락에서 환자 상태 변화를 반영해 순차적으로 최적 행동을 학습하는 강화학습(Reinforcement Learning, RL)이 주목받고 있으며, AI Clinician[2]을 비롯해 과거 임상 데이터를 활용한 다양한 모델이 제안되어왔다. 의료 분야에서는 환자를 대상으로 직접 치료 행동을 탐색하는 것이 윤리적, 안전성 문제로 불가능하기에 대부분의 연구는 기존 전자의무기록(Electronic Medical Record, EMR)을 활용한 오프라인 강화학습(Offline-RL)으로 진행된다. 그러나 기존의 Offline-RL 기반 패혈증 치료 연구는 주요 한계가 체계적으로 분석되지 않아, 임상 적용 시 안전성과 일반화를 위한 통합된 연구로 이어지기 어렵다. 이에 본 논문은 기존의 Offline-RL 기반 패혈증 치료 연구의 주요 한계를 체계적으로 분석하고, 연구의 재현성·안전성·일반화 관점에서 고려해야 할 체크리스트를 제안한다.

2. 사전지식: MDP와 Offline-RL

패혈증 치료와 같은 순차적 의사결정은 MDP(Markov Decision Process)로 모델링할 수 있다[3]. 이는 $M=(S,A,P,r,\gamma,\rho_0)$로 표현되며, 상태 $s\in S$는 환자 상황, 행동 $a\in A$는 가능한 치료 선택 범위를 규정하는 집합 공간이다. 에이전트는 보상 $r(s,a)$를 최대화하도록 순차 결정을 내리고, 치료에 대한 환자 반응의 불확실성을 상태 전이 $s' \sim P(\cdot|s,a)$로 기술한다. 또한, 초기 상태 분포 $\rho_0(s)=\Pr[s_0=s]$는 에피소드 시작 상태의 규칙을 제공한다. 정책 $\pi(a|s)$는 각 상태에서 행동을 선택하는 규칙이며, Offline-RL의 경우에는 과거 임상의의 실제 치료 결정을 따르는 행동정책 $\beta(a|s)$가 생성한 고정 데이터셋 $D=(s_i,a_i,r_i,s'_i)_{i=1}^{N}$만으로 정책을 학습한다. 이때 각 샘플의 상태 - 행동 쌍은 행동정책

하에서 형성된 상태 분포 d^{β}와 그 조건부 행동분포의 곱인 $(s_i,a_i) \sim d^{\beta}(s)\beta(a|s)$에서 나온다.

정책 학습의 목표는 장기 누적보상인 $J(\pi) = E\rho_0, P, \pi[\Sigma_{t=0}^{\infty}\gamma^t R_t]$을 최대화하는 최적 정책 π^*을 찾는 것이고, Offline-RL에서는 과거 정책보다 더 높은 장기 누적보상을 주는 학습정책을 찾는다. 이때 정책 π 하에서 상태 s에서의 기대 보상의 합인 상태-가치 함수는 $V^{\pi}(s) = E_{\pi}[\Sigma_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0 = s]$로 표현된다. 여기서 할인율 $\gamma \in [0,1)$은 현재 보상 대비 미래 보상의 중요도를 조절하여, 단기 안정과 장기 생존 사이의 균형을 정한다. 다음으로, 상태 s에서의 행동 a의 기대 보상의 합은 행동-가치 함수 $Q^{\pi}(s,a) = r(s,a) + \gamma E_{s' \sim P(\cdot|s,a)}[V^{\pi}(s')]$로 표현되며, 최적 정책을 따를 때 얻을 수 있는 최대 기대 보상은 $Q^*(s,a) = r(s,a) + \gamma E_{s' \sim P(\cdot|s,a)}[\max_{a'} Q^*(s',a')]$인 벨만 최적 방정식으로 표현된다. 결국, 정책 학습의 목표는 Q^*에 도달하는 정책 π^*를 찾는 것으로, MDP 설계는 성능과 임상 타당성에 직접적 영향을 주므로 정확하고 타당하게 설계하는 것이 중요하다.

3. 기존 Offline-RL 기반 패혈증 치료 연구의 한계

3.1. 데이터 기반 한계

- **품질·표본 제약**: ICU 데이터의 근복적인 한계는 센서·입력 오류로 인한 이상치, 불규칙 측정 간격으로 인한 상태 왜곡이며, 중증군 에피소드가 조기 사망/전원 등으로 짧게 검열되어 동일 상태-행동 반복 관측이 부족해 가치 추정 분산이 커질 수 있다. 한편, 변수 선택·정규화·스케일링 등 전처리의 작은 차이는 성능에 큰 영향을 끼치며, 에피소드 경계와 고정 1/2/4시간 등의 시간 이산화 방법은 시간 간격 간 결과 민감도가 크고, 연속 처치·약효 지연·누적 효과를 반영하지 못한다. 결측치의 경우 단순 대치/삭제하면 MNAR(Missing Not At Random)인 결측 패턴 자체의 임상 신호 손실 가능성이 존재한다.
- **분포 편향·일반화 취약성**: 단일 데이터셋으로 학습한 정책은 타 기관으로 전이 시 성능 저하가 잦으므로 다기관·다시기 학습과 교차 외부 검증이 필요하다. 특히, 패혈증 초기 결정이 이루어지는 응급실 환경은 ICU 데이터 기반 정책과 목표와 제약 조건이 불일치하여 성능이 떨어진다[4]. 처치의 경우, 임상의 관행·자원·환자 선호 등에 좌우되는 편향을 가질 수 있으며, 이는 의료 불평등을 키울 수 있음에 주의해야 한다.

3.2. MDP 설계 한계

- **상태 표현**: 일반적으로 선택되는 임상 변수 중심의 협소한 피처로는 고차원·비선형 병태를 포괄하기 어렵고, 군집 기반 축소는 시간 맥락을 포착하지 못한다[5]. 딥러닝 모델을 이용한 시계열 표상은 성능을 높이지만 해석가능성을 낮춘다. 한편, 미측정 교란(중증도, 검사 시행 결정, 결측 패턴 등)을 반영하지 못하면 최적 행동이 체계적으로 치우칠 수 있다.
- **행동 공간**: 관행적인 수액×바소프레서 5×5 이산화는 약물의 정밀 용량을 추천하지 못하며, 약물 종류·농도·투여 경로, 산소/환기/신대체요법 등을 누락해 임상 포괄성과 실행 가능성을 떨어뜨린다[6].
- **보상 설계**: 생존 여부와 같은 희소·지연 보상만으로는 학습이 불안정하여, 많은 연구에서 SOFA[6]·젖산 같은 중간보상을 도입했으나 성능 보장을 위해 보상 최적화 작업이 필요하다. 한편, 아직 부작용·자원·비용을 고려한 연구가 존재하지 않으며, 이로 인해 단기 지표 과최적화가 발생할 위험이 있다.

3.3. Offline-RL의 본질적 제약

Offline-RL은 탐색이 불가하여 데이터에 존재하지 않거나 희소한 상태-행동 영역인 OOD(Out-Of-Distribution)에 대해 Q 값을 신뢰성 있게 추정하기 어렵다. 그 결과 과대추정과 외삽 오류가 발생해 임상적으로 안전하지 않은 처방으로 이어질 수 있다. 또한, 혁신성-보수성의 딜레마가 존재하는데, 보수적 방법은 분포 밖 행동을 억제해 안전성은 높이지만 새로운 전략 발견을 제한하고, 반대로 보수성을 낮추면 극단 처방이 늘어 위험이 커진다. 또한, 정책 평가에서 OPE(Off-policy Evaluation)는 가정과 구현 세부에 민감하고 분산이 커 신뢰구간이 넓어질 수 있어, 다중 지표로 교차 검증이 필요하다.

3.4. 임상 적용성·안전·윤리·규제

임상 적용을 위해서는 안전·윤리·규제 차원의 고려가 필수다. 보상 극대화형 정책은 과격한 용량·처치를 제안할 수 있으므로 최대용량·금기·약물상호작용 등 가드레일을 미리 설정하고, 모델 불확실성을 인지한 의사결정 절차를 갖춰야 한다. 전문가 지식 주입이나 모방학습은 보완책이지만, 불완전한 지식을 그대로 학습하면 비최적의 수렴을 초래할 수 있다. 또한, 책임소재에 대한 논쟁, 단계적 임상시험 등 규제 정합성을 충족해야 실제 사용이 가능하다. 딥러닝 기반 정책의 해석가능성 부족은 추천 이유·

대안 제시를 어렵게 만들어 수용성을 낮추므로 설명·시각화 및 임상적 근거 제시가 필요하다.

4. 재현성·안전성·일반화를 위한 체크리스트

다음으로는 기존 Offline-RL 기반 패혈증 치료 연구가 가진 한계를 극복하기 위해, 연구의 재현성·안전성·일반화 관점에서 연구자와 임상가가 반드시 고려해야 할 핵심 체크리스트를 제안한다. 이는 데이터, MDP 설계, Offline-RL 평가, 안전성·설명가능성·공정성의 4가지 축을 중심으로 구성된다.

4.1. 데이터 측면

우선 <표 1>의 데이터 측면 체크리스트의 출발점으로, 다기관 데이터 사용 여부를 확인한다. 다기관 데이터 사용 시에는 이질성 정렬을 선행하고, 분포 차이를 정량 지표로 나타낸다. 이후, 학습 데이터의 미관측 영역이 큰 구간에서는 외삽 오류 위험이 있으므로 상태-행동 분포를 시각화할 필요가 있다. 다음으로 결측·이상·시간축을 점검한다. 결측은 우연보다 MNAR일 가능성이 있기에 단순 대치 대신 결측 패턴을 신호로 모델링하고, 이상치는 검출·완화 규칙을 명시한다. 마지막으로 하위군 공정성(성별·연령·인종 등)을 별도 보고하고, 교차 데이터셋으로 외부 검증을 거쳐 일반화 가능성을 확인하면 데이터 기인 편향과 외삽 위험을 줄일 수 있다.

<표 1> 데이터 측면에서의 체크리스트

Checklist Items	Check Questions
Use of multi-center data	Are multi-center (multi-site) data used?
Heterogeneity harmonization (if multi-center)	Have differences across sites, time periods, coding systems, and measurement units been harmonized?
Missing-data handling	Did you consider that missingness may be MNAR rather than MCAR, and handle it accordingly?
Data volume / effective sample size	Is the training dataset large enough, with sufficient coverage of state-action pairs, including the severe cohort?
Rare-action constraints / adjustment	Did you constrain or adjust for rare actions (e.g., regularization, caps, re-weighting)?
Distribution visualization & analysis	Have you visualized/analyzed the distribution of state-action pairs in the training data?
Irregular time-series handling	Have you addressed irregular sampling intervals (uneven measurement times)?
Distribution bias & subgroup fairness	Are there subgroup performance/fairness issues (e.g., by sex, age, race)?
Outlier handling	How are outliers detected and handled?

4.2. MDP 설계 측면

다음으로 <표 2>는 MDP 설계에서 발생하기 쉬운 시간 축 왜곡, 상태 표상 불충분, 축소된 행동 공간, 단선적 보상 문제를 선제적으로 차단하기 위한 점검 목록이다. 먼저 시간 해상도와 약효 지연을 검증하고, 시계열 임베딩과 임상 리뷰로 상태 표현의 타당성을 확인한다. 다음으로 행동 공간을 임상적으로 포괄하도록 확장하고, 보상에는 부작용·자원·비용을 포함해 단기-장기 목표의 균형을 맞춘다. 끝으로 임상의와의 명시적 합의와 가드레일을 문서화해, 정책 추천의 안전성과 수용성을 동시에 확보한다.

<표 2> MDP 설계 측면에서의 체크리스트

Checklist Items	Check Questions
Time Resolution	Have you performed sensitivity analysis for different resolutions (e.g., 1/2/4 hours)?
State Representation: Variable List	Have you specified the list of considered state variables?
State Representation: Complexity	Have you reflected temporal and nonlinear characteristics such as trends, variability, and interactions?
State Representation: Safety	Have you incorporated adverse, cumulative, and residual effects?
Action Space: Completeness	Does the action space include all core clinical treatments?
Action Space: Safety	Have you included safeguards to prevent overdosing?
Reward Design: Complexity	Have you included complications, side effects, resource constraints, and economic costs?
Reward Design: Safety	Are the clinical goals agreed upon with medical staff reflected?
Reward Design: Short-Long Balance	Have you reported the trade-offs between short-term stability and long-term outcomes?

4.3. Offline-RL 평가 측면

<표 3>에서는 Offline-RL의 평가 측면에서 점검사항을 제안한다. 외삽 오류를 방지하기 위해 여러 연구에서 CQL(Conservative Q-Learning)를 적용했듯[7],[8], 보수적 알고리즘으로 희귀 행동의 과대평가를 억제하되, 과도한 보수성으로 혁신적 전략 탐색이 차단되지 않도록 균형을 잡아야 한다. 희귀 구간에서는 불확실성을 정량적으로 보고한다. 정책의 안전성과 일반화를 확보하기 위해서는 최소 두 가지 이상의 OPE 기법을 적용하는 것을 권장하며, 교차 데이터셋 검증으로 일반화 가능성을 확인해야 한다.

<표 3> Offline-RL 평가 측면에서의 체크리스트

Checklist Items	Check Questions
Outlier Suppression	Have you suppressed out-of-distribution actions?
Uncertainty Reporting	Have you quantitatively reported uncertainty in state-action pairs?
Multiple OPE	Have you used at least two or more offline policy evaluation methods?
External Validation	Have you confirmed generalizability through validation with external datasets?

4.4. 안전성·설명가능성·공정성 측면

AI가 제시하는 추천이 임상적으로 납득되지 않는다면 의료진은 이를 신뢰할 수 없고, 결국 적용 자체가 불가능하다. 따라서 안전성·설명가능성·공정성을 모두 충족할 때 RL 정책은 실제 임상 적용과 규제 승인의 가능성을 확보할 수 있다. <표 4>에서는 이러한 관점에서의 점검 항목을 제안한다. 정책은 국제 가이드라인과 충돌하지 않아야 하며, 추천 행동의 분포가 실제 의사의 처방 패턴과 지나치게 괴리되지 않도록 관리되어야 한다. 또한, 사례 기반 설명이나 특징 중요도를 통해 추천 이유를 명확히 제시할 수 있어야 한다. 나아가 성별, 연령, 인종 등 환자 하위군별 성능을 분리 평가해 불균형적 위험을 방지하고, 정책이 의료진의 의도와 공정성을 함께 반영하는지 검증할 필요가 있다.

<표 4> 안전성·설명가능성·공정성 측면에서의 체크리스트

Checklist Items	Check Questions
Clinical Relevance & Explainability	Can the recommendation be justified through case-based explanation or highlighting specific importance?
Guideline Consistency	Does the recommendation conflict with international clinical guidelines?
Action Distribution Consistency	Does the distribution of recommended actions align with real-world physician practices?
Ethics & Fairness	Have you evaluated performance separately across subgroups (e.g., gender, age, race) and ensured no unfair risk occurs in specific subpopulations?
Policy Intention Validation	Have you validated the policy's alignment with medical intent and appropriateness using methods such as IRL(Inverse-RL)?

5. 결론

본 연구는 Offline-RL이 패혈증 치료의 복잡한 의사결정을 지원할 잠재력을 지니지만, 데이터, MDP 설계의 불완전성, Offline-RL의 탐색 불가, 그리고 안전성·일반화 검증 부족 등으로 임상 적용에 한계가 있음을 지적한다. 이는 단순 성능 문제가 아니라 실제 현장에서의 수용성과 안전성을 가로막는 핵심 장애물이다. 이에 본 연구에서는 재현성·안전성·임상 수용성을 높이기 위한 최소 점검 항목을 체크리스트로 제안했다. 이 체크리스트는 연구자와 임상의가 RL 기반 치료 전략을 개발하고 평가하는 과정에서 최소한의 안전장치이자 가이드라인 역할을 할 수 있으며, 향후 임상 적용 가능성을 높이는 기반이 될 것이다. 향후 연구는 의사, AI 연구자, 안전공학자, 윤리·규제 전문가가 함께 참여하는 운영 체계를 구축해 지속적으로 위험을 관리해야 할 것이다. 체크리스트의 각 항목의 최소 통과 기준에 대한 정량적인 전문가 합의를 마련함으로써 RL 기반 패혈증 치료가 연구 단계를 넘어, 실제 임상에서 활용 가능한 안전하고 일반화 가능한 도구로 발전하기를 기대한다.

사사문구

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 학·석사연계ICT핵심인재양성 지원을 받아 수행된 연구임 (IITP-2025-RS-2022-00156299)

참고문헌

[1]Singer, Mervyn, et al. "The third international consensus definitions for sepsis and septic shock (Sepsis-3)." Jama 315.8 (2016): 801-810.

[2]Komorowski, Matthieu, et al. "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care." Nature medicine 24.11 (2018): 1716-1720.

[3]Komorowski, Matthieu. "Clinical management of sepsis can be improved by artificial intelligence: yes." Intensive care medicine 46.2 (2020): 375-377.

[4]Nauka, Peter C., et al. "Challenges with reinforcement learning model transportability for sepsis treatment in emergency care." npj Digital Medicine 8.1 (2025): 1-5.

[5]Choi, Yunho, et al. "Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records." Communications Medicine 4.1 (2024): 245.

[6]Huang, Yong, Rui Cao, and Amir Rahmani. "Reinforcement learning for sepsis treatment: A continuous action space solution." Machine Learning for Healthcare Conference. PMLR, 2022.

[7]Tu, Rui, et al. "Offline Safe Reinforcement Learning for Sepsis Treatment: Tackling Variable-Length Episodes with Sparse Rewards." Human-Centric Intelligent Systems 5.1 (2025): 63-76.

[8]Nambiar, Mila, et al. "Deep offline reinforcement learning for real-world treatment optimization applications." Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. 2023.