의료 인공지능 리더보드 운영을 위한 정책 방향과 구현 방안

박준영[1], 김영재[2], 김성현[3], 신신애[3], 김광기[4]
[1]가천대학교 중개-임상 의학과
[2]가천대학교 길병원 가천의생명융합연구원
[3]한국지능정보사회진흥원 인공지능데이터본부
[4]가천대학교 의과대학 의공학교실
jun0613@gachon.ac.kr, Sashin@nia.or.kr, Kimcon@nia.or.kr, kimyj10528@gmail.com, kimkg@gahon.ac.kr

Policy Directions and Implementation Strategies of Medical AI Leaderboards

Jun Youn Park[1], Young Jae Kim[2], Sung Hyun Kim[3], ShinAe Shin[3], Kwang Gi Kim[4]
[1]Dept. of Translational-Clinical Medicine, Gachon University
[2]Gachon Biomedical & Convergence Institute, Gachon University Gil Medical Center
[3]Dept. of AI Data, National Information Society Agency
[4]Dept. of Biomedical Engineering, Pre-medical Course, College of Medicine, Gachon University

요 약

의료 인공지능(AI) 기술의 발전과 함께, 다양한 모델의 성능을 객관적으로 비교하고 검증할 수 있는 의료 리더보드의 필요성이 대두되고 있다. 본 연구는 의료 리더보드 운영을 위한 정책 방향과 구현 방안을 제안한다. 주요 내용으로는 의료 데이터의 특수성을 고려한 평가 지표 설계, 환자 안전성과 임상적 유효성을 담보할 수 있는 검증 절차, 개인정보 보호 및 비식별화 기준, 참여기관 간의 공정한 비교를 위한 표준화된 데이터셋 구축 등이 포함된다. 또한, 지속 가능한 리더보드 운영을 위한 제도적·기술적 지원체계와, 공공성과 투명성을 보장하는 거버넌스 모델에 대해 논의하였다. 본 연구는 향후 의료 AI 기술의 신뢰성과 활용도를 높이기 위한 리더보드 정책의 기본자료로 활용될 수 있을 것으로 기대된다.

1. 서론

의료 인공지능(AI) 기술은 진단, 예측, 판독 등 다양한 임상 영역에서 활용이 확산되고 있으며, 이에 따라 모델의 성능을 객관적으로 평가하고 비교할 수 있는 체계에 대한 필요성이 커지고 있다[1]. 특히, 의료 분야는 환자의 생명과 직결되는 높은 수준의 신뢰성과 안전성이 요구되기 때문에, 단순한 정확도 비교를 넘어서 임상적 유효성과 공정성을 반영한 평가 방식이 중요하다[2].

이러한 배경에서 의료 리더보드는 다양한 인공지능 모델을 표준화된 조건 하에 검증하고, 공정하게 성능을 비교·공개하는 수단으로 주목받고 있다[3]. 그러나 현재 운영되고 있는 대부분의 리더보드는 의료 데이터의 특수성과 민감성을 충분히 반영하지 못하고 있으며, 임상 실사용을 고려한 평가지표나 데이터 처리 기준도 미비한 실정이다[4]. Table 1은 의료 인공지능 리더보드 관련 현황 및 개선 방향에 대한 것이다.

본 연구는 이러한 한계를 극복하고자, 의료 리더보드 운영을 위한 정책 방향과 구현 방안을 제시함으로써, 향후 의료 AI 기술의 신뢰성과 임상 적용 가능성을 높이는 데 기여하고자 한다.

Table 1. 의료 인공지능 리더보드 관련 현황 및 개선 방향

AS-IS (현황)	TO-BE (개선 방향)
성능 비교 위주 및 정확도 중심	임상 유효성과 공정성 반영한 다면적 평가
비식별화·보안 기준 미흡	의료 데이터 특성 고려한 보호체계 확립
일반 AI 지표 사용 (Accuracy 등)	환자 안전성과 임상 맥락 고려한 지표 추가
기술 간 비교에 국한	인허가, 실증, 정책 연계 가능 구조로 확대

2. 의료 리더보드의 개념 및 역할

2.1. 리더보드 정의 및 의료 AI에서의 기능

리더보드는 인공지능(AI) 모델의 성능을 동일한 조건 하에서 평가하고, 그 결과를 순위 형태로 시각화하거나 공개하는 플랫폼을 의미한다. 일반적으로는 머신러닝 경진대회에서 많이 활용되며, 참가자들은 공통된 테스트 데이터셋에 대해 모델을 제출하고, 사전에 정의된 지표에 따라 점수를 부여받는다.

의료 AI 분야에서 리더보드는 단순한 성능 비교를 넘어 다음과 같은 기능을 수행한다. 첫 번째, 동일한 환자 데이터와 평가 지표를 기반으로, 다양한 알고리즘의 성능을 정량적으로 비교할 수 있도록 한다. 이는 알고리즘의 실제 임상 적용 가능성을 판단하는 데 기초 자료가 된다. 두 번째, 공개된 성능 순위를 통해 연구자 간 경쟁을 유도하고, 새로운 모델이나 구조에 대한 지속적인 연구와 발전을 장려한다[5]. 세 번째, 중앙화된 시스템에서 평가가 이루어지기 때문에, 특정 기관이나 기업이 평가 과정을 임의로 조작하기 어렵다. 이는 의료 AI에 대한 사회적 신뢰를 높이는 데 기여한다. 네 번째, 성능이 입증된 모델에 대해서는 추후 실증적 근거를 바탕으로 의료기기 인허가, 임상시험 연계, 실사용 검증 등의 후속 절차로 연결될 수 있다. 이처럼 의료 AI 리더보드는 기술적 성능을 넘어서, 임상 실효성, 공공성, 그리고 정책적 활용 가능성까지 고려한 평가 플랫폼으로서의 기능을 수행한다.

2.2. 기존 리더보드 사례 및 한계점 요약

의료 인공지능 분야에서는 다양한 공개 리더보드를 통해 모델 성능을 비교하고 연구 경쟁을 촉진하고 있다. 대표적인 사례로는 Grand Challenge, Medical Segmentation Decathlon (MSD), RSNA AI Challenges, 그리고 MedMNIST Benchmark 등이 있으며, 각 플랫폼은 의료 영상(CT, MRI, X-ray 등)을 기반으로 특정 진단 또는 분할 과제를 설정하고, 참가자 모델의 성능을 정량적으로 평가하여 순위를 제공한다. LLM 분야에서는 허깅페이스가 지속적으로 리더보드를 운영하고 있으며 의료 영상분야에서는 The Medical Image Computing and Computer Assisted Intervention Society (MICCAI)의 챌린지시리즈가 대표적인 러더보드이다. 2025년 MICCAI 챌린지에는 NIA의 AI 데이터 사업으로 구축된 병리데이터의 리더보드도 운영된다.

Table 2. 기존 리더보드 한계점 요약

주요 한계	설명
단순한 평가 지표	정확도 등 수치 중심으로 임상적 중요도 반영 부족
제한된 데이터 다양성	특정 기관/환경에 편중된 데이터 사용
현실과의 차이	실제 의료현장의 복잡한 상황 반영 어려움
과도한 제출 최적화	리더보드 점수만을 위한 과적합 우려

3. 운영을 위한 정책 방향

3.1. 평가 지표 설계 원칙

의료 인공지능 모델의 성능을 평가하기 위한 리더보드에서는, 단순한 수치 기반의 정확도만으로는 임상적으로 의미 있는 비교가 어렵다. 특히 의료 AI는 환자 진료와 직결되기 때문에, 평가 지표 또한 임상적 중요성과 해석 가능성, 비교 가능성을 고려하여 설계되어야 한다[6]. 임상적 중요도를 반영하는 지표가 포함되어야 한다. 일반적인 인공지능 평가에서는 정확도(accuracy), 정밀도(precision), F1-score 등이 사용되지만, 의료 영역에서는 민감도(sensitivity), 특이도(specificity), 양성예측도(PPV), 음성예측도(NPV) 등의 임상적 지표가 더 중요할 수 있다[7]. 또한, 복합적인 지표 활용이 권장된다. 단일 지표만으로는 모델의 성능을 포괄적으로 평가하기 어려우므로, AUROC, F1-score, Dice coefficient, clinical accuracy 등의 지표를 함께 제시하고 종합적인 판단이 가능하도록 해야 한다[8]. 비교 가능성과 재현성이 보장된 표준화된 평가 지표가 필요하다. 이를 통해 특정 모델이나 기관에 유리하게 작동하는 평가를 방지하고, 누구나 동일한 조건에서 모델 성능을 재현하고 비교할 수 있는 환경이 조성되어야 한다.

3.2. 환자 안전성 및 임상 유효성 반영

의료 인공지능 모델의 실제 임상 적용 가능성을 판단하기 위해서는, 단순한 기술적 성능을 넘어 환자 안전성(safety)과 임상 유효성(clinical utility)을 반영한 평가 체계가 필수적이다. 리더보드는 다양한 AI 모델을 동일한 조건에서 비교할 수 있는 도구로서 기능하지만, 평가 지표와 운영 기준이 임상 현실을 충분히 반영하지 못할 경우, 실제 환자 치료에 적용되기에는 한계가 있다.

환자 안전성 확보를 위해 평가 지표는 오류 발생 시 임상적 위험이 큰 지표, 예를 들어 false negative나 delayed detection에 대한 민감도를 반영해야 한다. 예를 들어 악성 질환의 조기 진단을 목적으로 하는 모델에서 민감도(sensitivity)는 단순한 정확도보다 더 중요한 기준이 될 수 있다. 또한 특정 질환에서는 오진(misdiagnosis)보다 미진단(missed diagnosis)이 더 치명적일 수 있음을 감안해야 한다.

평가에 사용되는 데이터셋도 환자 안전성과 유효성 확보에 중요한 역할을 한다. 훈련 및 테스트 데이터가 실제 임상에서 마주하는 다양한 상황(예: 기기 차이, 질환 중복, 이미지 품질 저하 등)을 포함하지 않을 경우, 높은 리더보드 점수에도 불구하고 실제 사용 시 성능 저하가 발생할 수 있다.

리더보드 설계 시에는 단순한 알고리즘 성능이 아닌, 환자 중심의 평가 기준과 임상 환경에서의 적합성이 반영되어야 한다.

3.3. 개인정보 보호 및 공정성 확보

의료 리더보드 운영에서 가장 중요한 전제 중 하나는 환자의 개인정보를 보호하는 동시에, 모든 참여자에게 공정한 비교 환경을 제공하는 것이다. 의료 인공지능 모델의 학습과 검증에는 고품질의 의료 영상 및 임상데이터가 필수적으로 요구되며, 이는 환자의 민감한 개인정보를 포함할 수밖에 없다. 따라서 리더보드 운영자는 기술적, 제도적, 윤리적 측면에서의 보호 장치를 함께 마련해야 한다[9].

모든 데이터는 비식별화 또는 가명처리 과정을 거쳐야 하며, 환자의 신원을 특정할 수 있는 모든 정보는 제거되어야 한다[10]. 또한 공정성 확보를 위해 모든 참가자는 동일한 조건의 테스트셋, 동일한 평가지표, 동일한 제출 규칙에 따라 평가를 받아야 한다.

4. 구현 방안 및 운영체계

4.1. 표준화된 데이터셋 구성

의료 리더보드의 평가 신뢰도를 확보하기 위해서는 표준화된 데이터셋을 구성하는 것이 필수적이다. 특히 의료 영상 및 임상 데이터를 기반으로 하는 AI 평가에서는 데이터의 품질, 다양성, 형식 일관성 등이 모델 성능에 직접적인 영향을 미치기 때문에, 체계적인 설계 원칙이 요구된다. Figure 1은 다양한 의료 데이터에 대한 표준화이다.

첫째, 형식적 표준화가 필요하다. 데이터 포맷(DICOM, PNG, CSV 등), 해상도, 채널 수, 메타데이터 구성 방식 등을 통일하여 모든 참여자 동일한 방식으로 데이터를 읽고 처리할 수 있도록 하는 것이다. 둘째, 내용적 표준화가 요구된다. 이는 질환의 정의 기준, 판독 기준, 라벨링 방식 등에서 임상 전문가와 협의된 프로토콜을 기반으로 데이터셋이 구성되어야 함을 의미한다. 셋째, 다양성과 대표성 확보가 중요하다. 리더보드의 평가가 실제 임상 적용 가능성과 연결되기 위해서는 단일 기관이나 특정 인구군에 편중되지 않은 다기관기 반의 데이터셋이 구성되어야 한다. 마지막으로, 표준화된 데이터셋은 평가용과 개발용으로 명확히 분리되어야 하며 테스트셋은 비공개 상태로 운영되어야 한다[11].

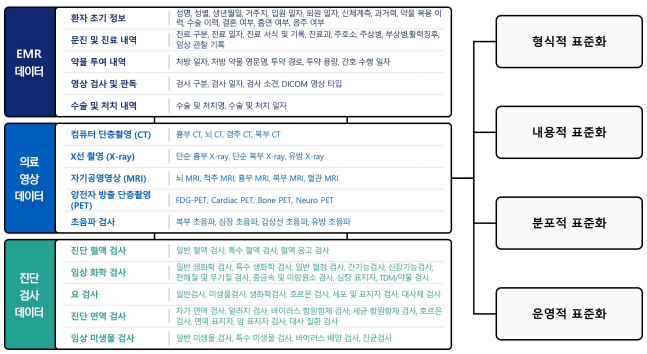

Figure 1. 다양한 의료 데이터에 대한 표준화 체계

4.2. 성능 검증 및 공개 방식

의료 리더보드의 핵심 기능은 다양한 인공지능 모델의 성능을 객관적으로 평가하고 그 결과를 공정하게 공개하는 것이다. 이를 위해서는 신뢰성 있는 성능 검증 절차와 투명한 결과 공개 방식이 함께 마련되어야 한다.

이러한 설계에서는 다음과 항목을 포함할 필요가 있다. 성능 검증 절차는 자동화되면서도 공정해야 한다. 참가자는 사전에 지정된 형식의 모델 출력(분류 결과, 세그멘테이션 마스크, 캡션 텍스트 등)을 제출하고 서버 측에서 블라인드 테스트셋에 대한 평가가 자동으로 이루어지는 구조가 비교적 가장 명확하다. 평가지표는 의료 분야의 특성을 반영하는 민감도(sensitivity), 특이도(specificity), AUROC, Dice coefficient, ROUGE 등 복합 지표를 포함해야 한다[12]. 모델 제출 및 평가 방식은 남용을 방지할 수 있도록 제한되어야 한다. 각 모델의 임상적 장단점, 대상 인구 집단별 성능 차이, 오류 유형 분석 등을 함께 제공함으로써 사용자의 이해를 돕고 활용도를 높일 수 있다[13].

4.3. 지속 가능성과 제도적 지원

지속 가능성을 위해 일회성 대회가 아닌, AI 성능을 지속 검증하는 장기적 플랫폼이 필요하다. 리더보드가 신뢰성과 활용도를 유지하기 위해서는 데이터셋의 정기적인 업데이트, 평가 기준 개선, 평가 환경 유지·보수 등이 필수적이다. 특히, 의료 환경은 새로운 질환, 임상 프로토콜의 변화가 빈번하기 때문에 이를 반영한 버전 관리 및 장기적 유지보수 계획이 필요하다. Figure 2는 지속 가능한 의료AI 리더보드 운영 체계이다. 리더보드 운영이 국가 인증, 인허가 심사, 연구 과제 평가 등과 연계될 수 있는 법적·제도적 장치가 마련되어야 한다[14].

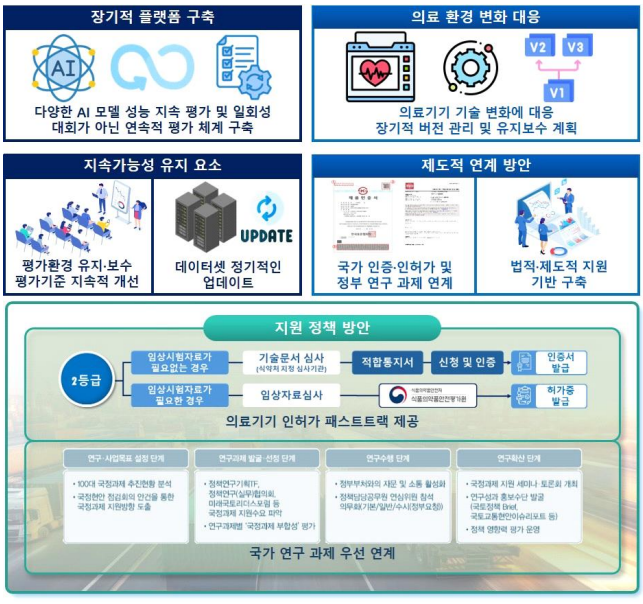

Figure 2. 지속 가능한 의료AI 리더보드 운영 체계

5. 결론

본 연구는 의료 인공지능 기술이 다양한 임상 영역에서 활용됨에 따라, 그 성능을 객관적이고 임상적으로 신뢰할 수 있는 방식으로 검증하고 비교할 수 있는 의료 리더보드의 필요성을 강조하였다. 리더보드는 단순한 알고리즘 성능 비교를 넘어서, 환자 안전성과 임상적 효용성을 함께 고려한 공공적 평가 플랫폼으로 기능해야 하며, 이를 실현하기 위한 정책적 방향성과 구현 전략이 병행되어야 한다.

의료 리더보드 운영을 위한 핵심 구성 요소로, 형식적·내용적·분포적·운영적 표준화에 기반한 데이터셋 구성, 임상 중심의 복합 평가 지표 설계, 자동화된 검증 절차, 공정성과 재현성을 보장하는 평가 방식, 그리고 결과의 해석 가능성과 투명한 공개 체계를 제시하였다. 특히 의료 환경의 변화 속도를 반영할 수 있도록 리더보드의 평가 기준과 데이터셋은 정기적으로 업데이트되어야 하며, 테스트셋은 블라인드 상태로 운영하여 과적합을 방지하고 모델의 일반화 능력을 평가할 수 있어야 한다.

특히 의료 환경의 변화 속도를 반영할 수 있도록 리더보드의 평가 기준과 데이터셋은 정기적으로 업데이트되어야 하며, 새로운 질환군의 등장, 영상장비 및 진단기술의 발전, 임상 프로토콜의 변경 등이 주기적으로 반영될 수 있는 유연한 데이터 관리 체계가 요구된다. 이 과정에서 단순히 기존 데이터를 대체하거나 추가하는 것이 아니라, 버전별 이력 관리와 함께 모델 성능의 추세를 비교할 수 있도록 설계되어야 하며, 동일 모델의 다양한 시점별 평가 결과를 축적·분석할 수 있는 시스템이 필요하다. 또한 테스트셋은 블라인드 상태로 운영하여 참가자에게 사전 노출되지 않도록 함으로써 과적합을 방지하고, 모델의 일반화 능력을 정직하게 평가할 수 있어야 한다. 이를 위해 사전 정의된 기준에 따라 구성된 검증 전용 서버 또는 제출-응답 기반 자동 평가 시스템이 도입될 수 있다.

아울러 리더보드의 지속 가능성과 제도적 기반을 확보하기 위해서는, 정부, 학회, 병원, 산업체 등 다양한 이해관계자들이 참여하는 협력 구조가 필요하며, 장기적 운영을 위한 독립적인 운영 주체와 재정적·행정적 지원 체계가 마련되어야 한다. 리더보드가 의료기기 인허가, 국가 과제 선정, 임상시험 설계 등과 정책적으로 연계될 수 있는 기반을 갖추는 것도 중요하다. 또한 우수한 성능을 보인 모델에 대해서는 의료 현장 실증 기회, 규제 완화, 인증 가속화 등 실질적인 인센티브를 제공함으로써 참여자 유인을 강화해야 한다.

향후 의료 리더보드는 의료 인공지능의 검증뿐 아니라, 신뢰할 수 있는 기술을 의료 현장에 안전하게 도입하기 위한 핵심 플랫폼으로 발전할 수 있다. 본 연구는 그러한 발전 방향을 구조화하여 제시하였으며, 향후 의료 AI 기술의 공공적 검증 체계를 구축하고 제도화하는 데 있어 기초 자료로 활용될 수 있을 것으로 기대된다.

References

[1] Cabitza, Federico, Raffaele Rasoini, and Gian Franco Gensini. “Unintended consequences of machine learning in medicine.“ Jama 318.6 (2017): 517-518.

[2] Topol, Eric J. “High-performance medicine: the convergence of human and artificial intelligence.“ Nature medicine 25.1 (2019): 44-56.

[3] Esteva, Andre, et al. “A guide to deep learning in healthcare.“ Nature medicine 25.1 (2019): 24-29.

[4] Maier-Hein, Lena, and Bjoern Menze. “Metrics reloaded: Pitfalls and recommendations for image analysis validation.“ arXiv. org 2206.01653 (2022).

[5] Willemink, Martin J., et al. “Preparing medical imaging data for machine learning.“ Radiology 295.1 (2020): 4-15.

[6] Kelly, Christopher J., et al. “Key challenges for delivering clinical impact with artificial intelligence.“ BMC medicine 17 (2019): 1-9.

[7] Liu, Xiaoxuan, et al. “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis.“ The lancet digital health 1.6 (2019): e271-e297.

[8] Raghu, Maithra, and Eric Schmidt. “A survey of deep learning for scientific discovery.“ arXiv preprint arXiv:2003.11755 (2020).

[9] Guerra-Manzanares, Alejandro, et al. “Privacy-preserving machine learning for healthcare: open challenges and future perspectives.“ International Workshop on Trustworthy Machine Learning for Healthcare. Cham: Springer Nature Switzerland, 2023.

[10] Ali, Mansoor, et al. “Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey.“ IEEE journal of biomedical and health informatics 27.2 (2022): 778-789.

[11] Chai, Yuan, et al. “Clinical benchmark dataset for AI accuracy analysis: Quantifying radiographic annotation of pelvic tilt.“ Scientific Data 11.1 (2024): 1162.

[12] Van Calster, Ben, et al. “Performance evaluation of predictive AI models to support medical decisions: Overview and guidance.“ arXiv preprint arXiv:2412.10288 (2024).

[13] Chai, Yuan, et al. “Clinical benchmark dataset for AI accuracy analysis: Quantifying radiographic annotation of pelvic tilt.“ Scientific Data 11.1 (2024): 1162.

[14] Hailemariam, Maji, et al. “Evidence-based intervention sustainability strategies: a systematic review.“ Implementation Science 14 (2019): 1-12.