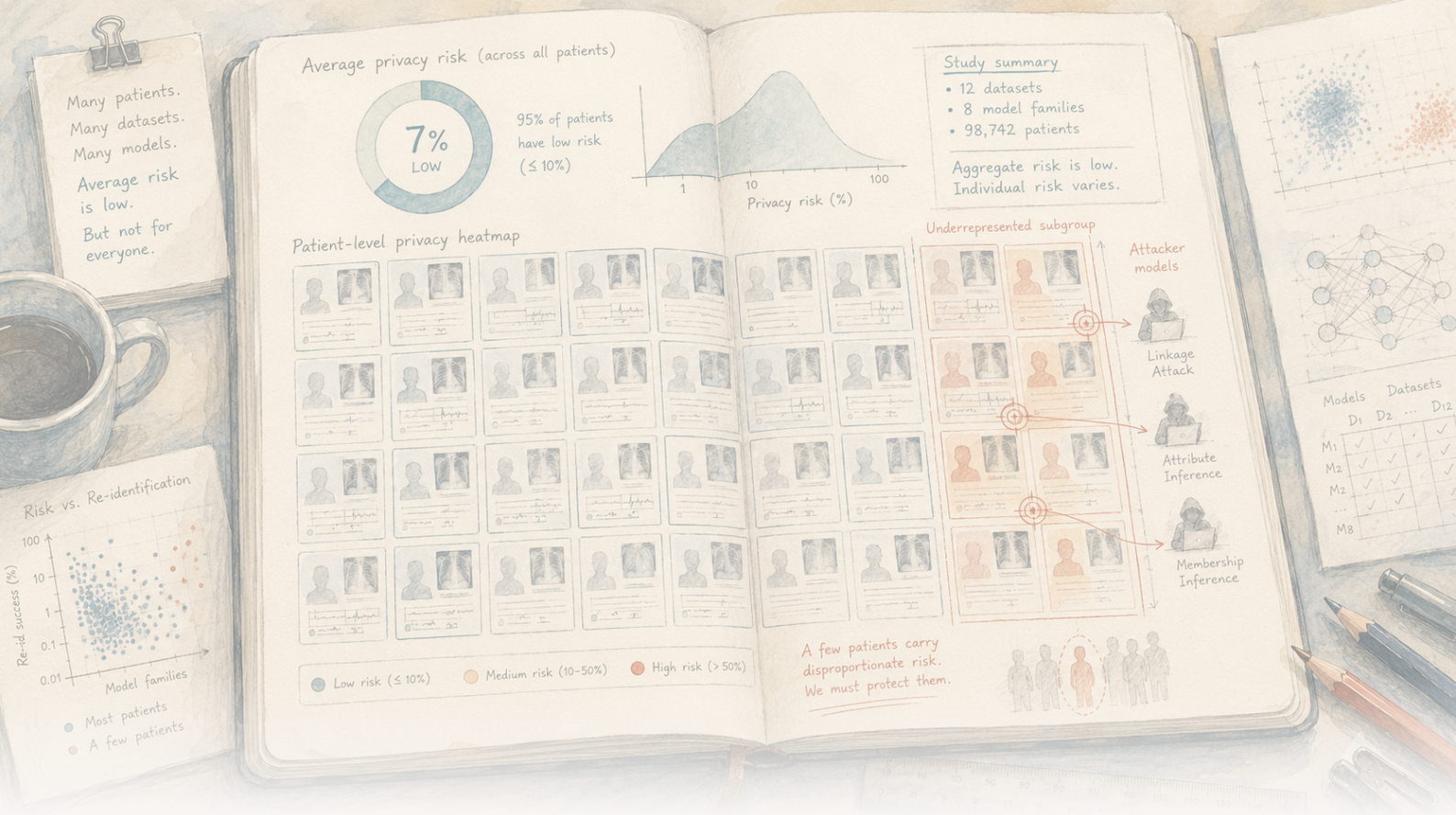


The CLEAN PAPER

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

의료 AI 는 평균적으로 프라이빗해도 특정 환자를 노출할 수 있다—특히 과소대표된 환자를

'프라이버시 보존형'의료 *AI* 라는 평가는 보통 평균 숫자 하나에 기대지만, 이 연구는 그것이 잘못된 시험이라고 주장한다. 특정 개인의 기록이 학습 데이터에 들어 있었는지를 알아내는 멤버십 추론 공격을 *7* 개 의료 데이터셋 *(*영상, *ECG,* 전자의무기록*)* 과 많은 모델에서 환자별로 측정한 결과, 평균적으로 안전해 보이는 모델에서도 일부 개인은 거의 완벽하게 식별될 수 있었다 *(*공격 $AUC \geq 0.95$*)*. 노출은 소수 인종·민족, 희귀질환, 특이 영상 등 과소대표 집단에 체계적으로 집중됐고 모델이 커질수록 악화됐다. 저자들은 의료 *AI* 를 포기하라고 하지 않는다. 환자별 프라이버시 평가, 모델 접근 통제, 차등 프라이버시를 요구한다. 불편한 핵심은'평균적으로 프라이빗하다'는 것이 프라이버시 보장이 아니라는 것이다.

Written by Lucio Vaglio · Cross-checked by Michele Renda · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Disparate privacy risks from medical AI*, Moritz Knolle, Georg Kaissis, Daniel Rückert and colleagues (Technical University of Munich; Imperial College London; Hasso Plattner Institute), Nature (2026) · DOI: 10.1038/s41586-026-10688-0

프라이버시 보장은 평균이 아니라 한 사람에게 하는 약속이다

의료 AI 모델을' 프라이버시 보존형' 이라고 부를 때 그 주장은 보통 숫자 하나에 기대고 있다. 모델을 학습한 모든 환자를 놓고 평균을 내면, 특정 개인이 학습 데이터에 들어 있었는지 추론할 확률이 낮다는 것이다. 안심되는 말처럼 들린다. 이 논문이 조용하지만 불편하게 지적하는 핵심은 **그 숫자가 잘못된 시험이라는 것**이다.

프라이버시는 평균이 아니다. 각 개인에게 하는 약속이다. 학습 세트에 포함됐다는 사실이 나중에 그 사람을 노출시키지 않을 것이라는 약속이다. 평균값은 거의 모든 사람에게 그 약속을 지키면서 소수에게는 완전히 깨뜨릴 수 있다. 연구진은 이전보다 훨씬 세밀하게 환자 한 명씩 프라이버시 위험을 측정했고 정확히 그런 현상을 발견했다. 전체 평균으로는 안전해 보이는 모델도 특정 개인이 학습 데이터에 들어 있었는지를 거의 완벽하게 누설할 수 있었고, 그렇게 노출되는 사람은 불균형하게도 이미 가장 덜 보호받는 집단에 속했다.

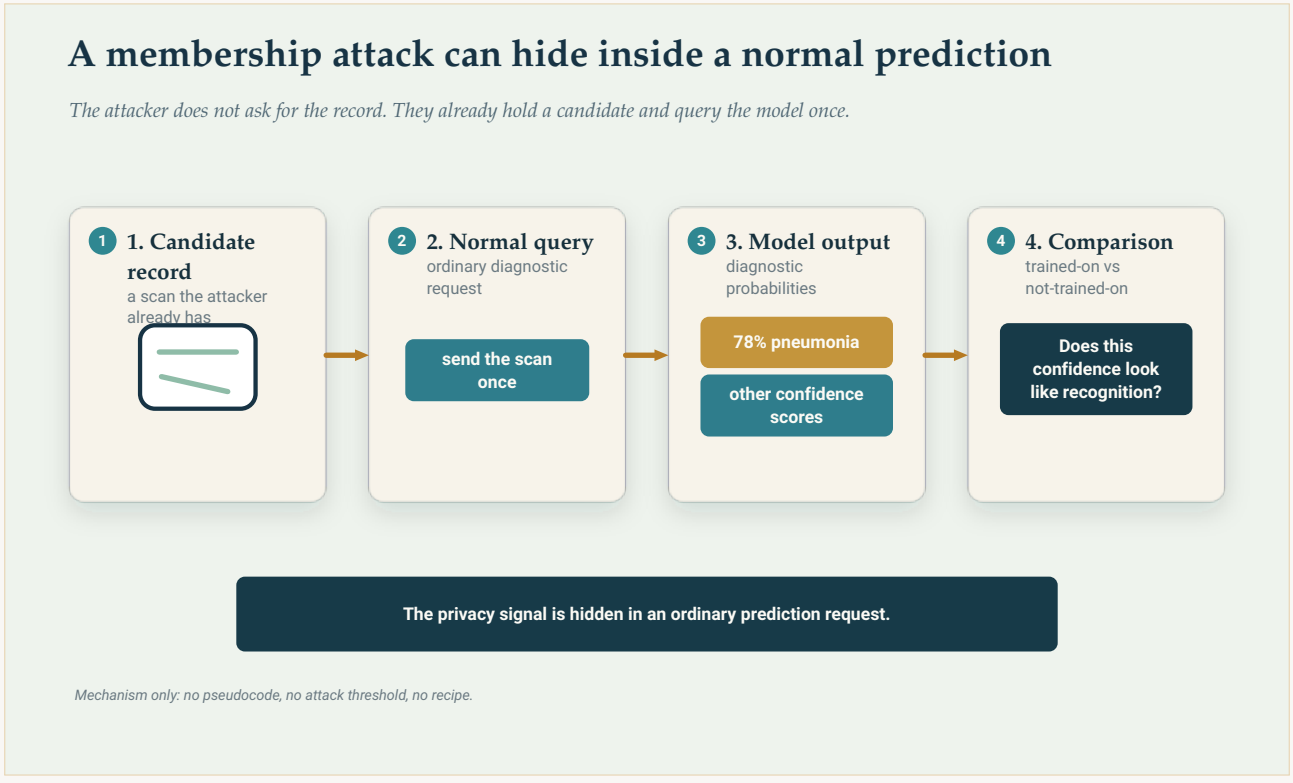
연구진이 한 일

Moritz Knolle 가 이끌고 뮌헨공대의 Daniel Rückert, Hasso Plattner Institute 의 Georg Kaissis, Imperial College London 동료들이 참여한 연구팀은 잘 알려진 프라이버시 위협인 **학습 데이터 포함 여부 추론 공격 (membership inference attack)** 의 질문을 바꿨다. ' 데이터셋 전체에서 평균적으로 이 공격은 얼마나 자주 성공하는가?' 가 아니라' 이 특정 환자는 얼마나 노출돼 있는가?' 를 물었다.

연구진은 의료영상, 심전도, 전자의무기록 등 성격이 매우 다른 **7 개의 확립된 의료 데이터셋**에서 환자별 분석을 수행했고, 데이터셋마다 200 개 모델을 다뤘다. 각 모델의 각 환자에 대해 공격자가 그 사람의 기록이 학습 데이터에 있었는지를 얼마나 확실하게 알아낼 수 있는지 추정한 다음, 질병 상태, 자기보고 인종, 보험, 성별, 영상 촬영 프로토콜 등 집단별로 결과를 나눠 보았다.

> **학습 데이터 포함 여부 추론 공격은 실제로 무엇인가?**
>
> 여기서의 누출은' 모델이 당신의 기록을 그대로 뱉는다' 보다 미묘하다. 즉, **당신의 데이터가 학습 세트에 들어 있었는지 여부** 자체가 새어 나오는 문제다.
> 먼저 이런 모델의 정상적인 사용을 생각해 보자. 흉부 X 선 같은 영상을 넣으면 폐렴 확률 78% 와 함께 심비대, 부종, 경결 등의 확률을 돌려준다. 공격은 바로 이 평범한 진단 응답만 사용한다. 특별한 비밀 인터페이스가 필요한 것이 아니다.
> 약점은 보통 모델이 한 번도 보지 않은 예시보다 **학습에서 정확히 봤던 예시를 조금 더 자신 있게** 판단한다는 데 있다. 시험 문제를 미리 본 학생을 떠올리면 된다. 연습했던 문제에서는 답이 조금 지나치게 빠르고 확신에 차게 나온다. 그 흔적을 보고 특정 기록이 모델의 학습 예시에 포함됐는지 판단하는 것이 이 공격의 핵심이다.
> 공격의 논리는 이렇다. 누군가 특정인의 영상이 모델 학습에 사용됐는지 알고 싶다고 하자. 모델에게 그 기록을 내놓으라고 요구하는 것이 아니다. 이미 후보 영상, 즉 그 사람의 실제 영상이나 매우 가까운 사본을 갖고 있다. 평범한 진단 요청처럼 한 번 넣고 모델의 확신도를 기록한다. 그 다음 이

확신도가 해당 영상을 학습한 모델의 반응에 더 가까운지, 학습하지 않은 모델의 반응에 더 가까운지 비교한다. 공격자는 특별한 장비 없이 일반 컴퓨터에서 자기' 참조 모델 (reference model)' 을 학습해 이런 과도한 자신감의 패턴을 익힐 수 있다. 실제 모델이 이 영상에 유난히 확신에 차 있다면, 마치 알아보는 것처럼 반응한다면, 그 영상이 학습 데이터에 있었다는 강한 증거가 된다.
불편한 점은 요청 자체가 공격처럼 보이지 않는다는 것이다. 임상의가 하는 것과 같은 진단 질의이고, 프라이버시 신호는 평범한 예측 안에 숨는다. 그래서 위험이 구체적이다. 다만 현재까지는 명시된 실험실 조건에서 입증된 위험이지 실제 현장에서 적발된 공격 빈도를 말하는 것은 아니다.
기록 내용 자체가 나오지 않는데도 포함 여부가 왜 중요한가? **포함 여부 자체가 정보**이기 때문이다. 어떤 모델이 특정 암 면역치료를 받은 환자들로 학습됐다면, 당신의 기록이 그 데이터에 포함됐다는 사실을 확인하는 것만으로도 당신이 그 암을 앓았을 가능성을 드러낼 수 있다. 보험사나 고용주가 알아서는 안 될 종류의 정보다. 내용은 봉인돼 있어도' 그 집단에 속했다' 는 사실이 빠져나온다.
저자들은 모델의 일반 예측에 접근할 수 있음, 후보 기록을 갖고 있음, 공격자가 자기 참조 모델을 만들 수 있음이라는 가정을 명확히 적는다. 공격 사용법을 알려주기 위해서가 아니라 위협을 막연히 취급하지 않고 정확히 논의하기 위해서다.

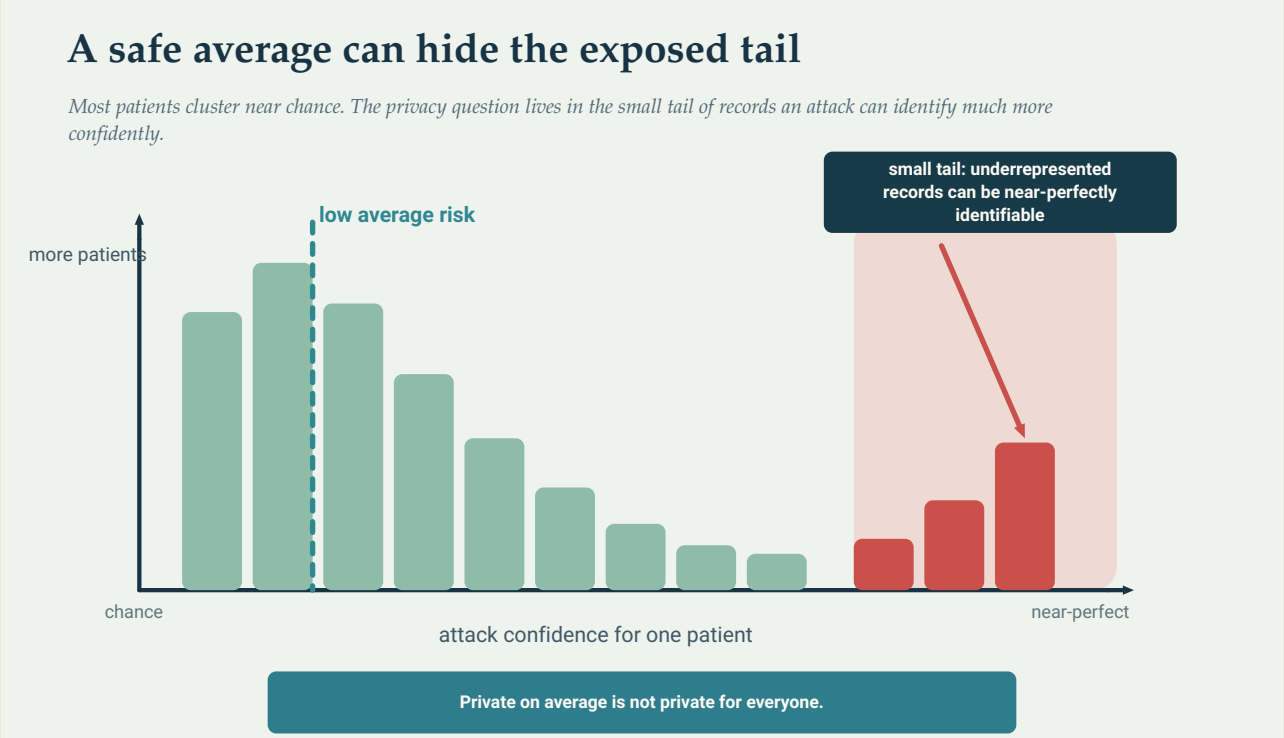


이 공격에는 특별한 프라이버시 API 가 필요하지 않다. 모델이 전에 본 기록에 유난히 높은 확신을 보인다면 평범한 진단 질의에서도 학습 데이터 포함 여부를 드러내는 신호가 새어 나올 수 있다.

무엇을 발견했나

세 가지 결과가 있으며 뒤로 갈수록 더 날카롭다.

평균은 노출된 사람을 숨긴다. 일반적인 방식대로 전체 평균을 내면 많은 모델이 꽤 안전해 보인다. 대부분 환자에서는 공격 성능이 우연보다 나을 것이 없다. 그러나 환자별로 보면 그림이 달라졌다. Knolle 는 연구 발표에서 이전 평가는' 항상 모든 환자의 평균 위험만 측정했다' 며' 처음으로 개별 환자 수준의 위험을 조사했더니 매우 다른 그림이 나왔다' 고 요약했다. 일부 개인에서는 공격이 **거의 완벽하게** 성공했다. 저자들이 측정한 공격 AUC 는 **0.95 이상**이었다 (0.5 는 동전 던지기 수준, 1.0 은 완벽한 구분). 데이터셋 전체 평균은 우연 수준처럼 보이는데도 그랬다.



평균 공격 성공률은 우연 수준에 가까울 수 있지만, 오른쪽의 작은 꼬리에는 훨씬 더 쉽게 식별되는 기록이 남는다. 논문의 핵심은 프라이버시를 전체 평균뿐 아니라 환자별로 확인해야 한다는 것이다.

노출은 불평등하며 대표성이 낮은 집단에 집중된다. 거의 완벽한 공격에 가장 취약한 환자는 데이터에서 **과소대표된 집단**에 체계적으로 속했다. 소수 인종·민족, 희귀질환 표현형, 비정상적 영상 특성이 그 예다. 전자의무기록 데이터셋에서는 가장 취약한 기록 가운데 흑인 환자가 예상보다 **31% 더 자주** 나타났고, 유방촬영 데이터에서는 악성 의심으로 표시된 영상이 극고위험 구간에서 **+1,179%** 나 과대표집됐다. (이 두 숫자는 전체 그림의 예시다. 논문은 여러 집단에서 같은 방향의 왜곡을 보고한다. 또한 이는 작은 극단 위험 꼬리 안에서의 상대적 과대표현이다. 즉 가장 노출된 기록 중 이 환자들이 예상보다 얼마나 자주 나타나는지를 뜻하지, 흑인 환자나 의심 유방촬영 환자 중 몇 퍼센트가 노출됐다는 뜻이 아니다.) 모델 안에 비슷한 사례가 적으면 해당 기록은 다른 기록들 속에 섞여 숨기 어렵고 더 튄다. 그리고 바로 그' 튀는 정도' 를 포함 여부 추론 공격이 감지한다. 프라이버시 실패는 무작위로 흩어지는 것이 아니라, 이미 데이터에서 주변부에 놓인 사람에게 집중된다.

모델이 커질수록 더 나빠진다. 거의 완벽한 공격에 노출되는 환자 수는 모델 용량과 함께 급격히 늘었다. 한 피부과 데이터셋에서는 가장 작은 모델에서 사실상 0 이던 비율이 가장 큰 모델에서 약 **10 명 중 1 명**으로 올라갔다. 의료 AI 가 더 크고 강력해지는 방향으로 발전하는 상황에서, 저자들은 이 특정 위험이 줄어드는 것이 아니라 심해진다고 경고한다.

Rückert 의 요약은 단호하다. '용납할 수 있는 위험이 아니다. 건강 데이터는 매우 민감하다.'

왜' 평균적으로 안전하다' 는 잘못된 시험인가

낮은 평균 위험을 '이상 없음' 판정으로 읽고 싶은 유혹이 있다. 논문의 핵심 교훈은 여기서 평균이 단순히 부정확한 것이 아니라 **잘못된 것을 측정하고 있다**는 점이다.

프라이버시 보장은 중간에 있는 사람이 아니라 가장 위험한 사람에게도 성립해야 의미가 있다. 1,000 명 중 999 명은 전혀 식별할 수 없지만 한 명은 거의 완벽하게 식별되는 모델을 그 한 사람에게 의미 있는 방식으로'99.9% 프라이빗' 하다고 부를 수는 없다. 게다가 그 한 사람이 예측 가능하게 희귀질환 환자나 소수 인종 환자라면 지표는 단순히 불완전한 것을 넘어 조용히 차별적이다. 다수를 위한 안전을 측정하고 모두를 위한 안전이라고 부르기 때문이다.

이 논문이 요구하는 질문의 변화는 명확하다. 이 모델은 평균적으로 얼마나 프라이빗한가? 에서 가장 노출된 환자는 누구이며, 그 사람들은 어떤 집단인가? 로 옮겨가야 한다. 두 질문은 다르고, 프라이버시 약속을 평가하는 데 중요한 것은 두 번째다.

이 연구가 증명하거나 주장하지 않는 것

- 의료 AI 를 포기해야 한다고 말하지 **않는다**. 저자들의 프레임은 후퇴가 아니라 완화다. 위험을 측정하고 고치자는 것이지 개발을 멈추자는 것이 아니다.
- 모든 의료 모델이 누출 중이거나 특정 배치 모델이 실제 공격을 받았다는 뜻이 **아니다**. 위험이 존재하고 불평등하게 분포하며 표준 전체 지표가 이를 놓친다는 것을 보여준다.
- 당신의 기록이 이미 노출돼 있다는 뜻이 **아니다**. 공격에는 모델 접근권, 후보 기록, 공격자 쪽 인프라라는 특정 조건이 필요하다. 아무나 즉흥적으로 할 수 있다는 주장이 아니다.
- 환자 기록의 내용이 새어 나온다는 것을 보여주지 **않는다**. 누출되는 것은 기록의 내용이 아니라 **학습 데이터에 포함됐는지 여부**다. 기록 전체가 출력되기 때문이 아니라 다른 이유로 위험하다.
- 격차를 대표 숫자 하나로 **축약하지 않는다**. 강하고 재현되는 결과는 하나의 수치가 아니라 패턴이다. 전체 지표가 개인 위험을 낮게 보이며, 과소대표 환자가 그 위험을 불균형하게 떠안는 현상이 여러 데이터셋과 수백 개 모델에서 반복된다.

근거는 얼마나 탄탄한가?

이는 경험적·방법론적 결과이고 상당히 강건하다. 전체 프라이버시 지표가 환자별 위험을 체계적으로 낮게 평가하고, 남은 위험이 과소대표 집단에 집중되며, 모델 용량이 커질수록 악화된다는 패턴이 **서로 다른 데이터 유형의 7 개 데이터셋**과 데이터셋별 많은 모델에서 유지됐다. 바로 이런 폭이 측정 주장에 신뢰성을 준다. 하나의 데이터셋에서만 나온 우연한 인공물이 아니다.

솔직한 주의점은 두 가지다. 첫째, 저자들이 만든 공격을 실제로 실행해 측정한 위험의 시연이다. 명시된 가정 아래 유능한 공격자가 얼마나 잘할 수 있는지를 보여주지, 현실 세계에서 공격이 얼마나 자주 일어나는지는 측정하지 않는다. 둘째, 일부 환자에서 거의 완벽한 공격 성공률이 나온다는 눈에 띄는 숫자는 설계상 가장 불리한 개인들을 묘사한다. 바로 그것을 보기 위한 연구이므로' 평균이 암시하는 것보다 꼬리가 훨씬 두껍다' 로 읽어야 하지' 대부분 환자가 노출돼 있다' 로 읽으면 안 된다.

적절한 태도는 공포도 무시도 아니다. 의료 AI 의 프라이버시를 인증하는 표준 방식이 자기 최악의 사례를 보지 못하고 있으며, 그 최악의 사례가 이미 여유가 가장 적은 환자에게 집중된다는 것을 여러 데이터셋에서 신중하게 보여준 연구다.

왜 중요한가

이 결과를 단순한 기술적 각주 이상으로 만드는 이유는 두 가지다.

첫째, ' 프라이버시 보존' 이라는 표현이 허용되려면 무엇을 의미해야 하는지를 바꾼다. 모델을 평균 프라이버시 점수와 함께 공개했을 때 그 평균이 정말 낮더라도, 포함 여부 추론 공격으로 거의 완벽하게 식별 가능한 환자 소집단을 숨길 수 있다. 논문의 실무적 요구는 구체적이다. 공개 전에 프라이버시 위험을 **환자별로** 평가하고, 배치된 모델에 누가 접근할 수 있는지 통제하며, **차등 프라이버시 (differential privacy)** 같은 기법을 사용하라는 것이다. 이는 학습 중 작고 정교하게 조절된 노이즈를 넣어 포함 여부 추론 공격을 어렵게 만들며, 모델 유용성에는 실제 비용이 따른다. 프라이버시-유용성 절충은 분명하고 공짜가 아니다. ' 평균을 확인했다' 는 말만으로는 더 이상 확인을 마쳤다고 해서는 안 된다는 주장이다.

둘째, 프라이버시와 공정성을 묶는다. 의료 데이터에서 과소대표돼 이미 의료 AI 의 서비스를 더 나쁘게 받을 가능성이 높은 바로 그 집단이, AI 가 프라이버시도 가장 약하게 보호하는 집단으로 나타났다. 첫 번째 불평등을 해결하려 애쓰는 분야가 두 번째를 남의 부서 문제로 취급할 수 없다. 같은 사람들이기 때문이다.

' 측정했고 평균 위험이 낮으니 괜찮다' 는 의료 AI 프라이버시의 안심 서사를 이 논문은 해체한다. 기술을 두려워하게 만들기 위해서가 아니라 기준을 제자리로 옮기기 위해서다. 가장 다칠 가능성이 높은 사람에게도 실제로 확인했을 때만 지켰다고 말할 수 있는 약속이어야 한다.

핵심 요약

학습 데이터 포함 여부 추론 공격은 특정 개인의 기록이 AI 모델의 학습에 사용됐는지를 알아내려 한다. 포함 여부 자체가, 예를 들어 특정 질환을 앓았다는 사실을 드러낼 수 있기 때문에 원래 기록 내용이 나오지 않아도 진짜 프라이버시 누출이 될 수 있다. 이전 연구는 데이터셋 전체에서 이런 공격이 평균적으로 얼마

나 성공하는지를 측정했다. 이번 연구는 의료영상, ECG, 전자의무기록을 포함한 7 개 의료 데이터셋과 많은 모델에서 환자별 위험을 측정했다. 그 결과 전체 지표는 위험을 크게 낮게 평가했다. 평균적으로 안전해 보이는 경우에도 일부 개인은 거의 완벽하게 식별될 수 있었고 (공격 AUC $\geq$ 0.95), 가장 취약한 사람은 소수 인종·민족, 희귀질환, 특이 영상 같은 과소대표 집단에 체계적으로 속했으며 모델이 커질수록 문제가 심해졌다. 저자들은 의료 AI 를 포기하라고 하지 않는다. 개인 수준에서 프라이버시를 측정하고, 모델 접근을 통제하며, 차등 프라이버시를 사용하라고 주장한다. 핵심은' 평균적으로 프라이빗하다' 는 것이 프라이버시 보장이 아니며, 그 평균이 실패하는 사람은 이미 가장 덜 보호받는 환자일 가능성이 높다는 것이다.

과장 없이 보면

논문이 보여주는 것: 7 개 의료 데이터셋과 각각의 많은 모델에서 일부 개인의 환자별 포함 여부 추론 위험은 전체 지표가 암시하는 것보다 훨씬 높고, 경우에 따라 거의 완벽한 공격 성공 (AUC $\geq$ 0.95) 에 이른다. 남은 위험은 과소대표 집단에 불균형하게 집중되고 모델 용량이 커질수록 증가한다.

그럴듯하지만 입증되지 않은 것: 현실의 공격자가 이미 배치된 임상 모델을 상대로 이를 악용하고 있다는 것. 연구는 명시된 가정 아래 공격의 능력과 위험 분포를 보여주지만 실제 공격 빈도를 측정하지 않는다.

보여주지 않는 것: 의료 AI 를 포기해야 한다는 것, 모든 모델이 누출하거나 특정 배치 모델이 침해됐다는 것, 환자 기록의 내용 자체가 노출된다는 것, 격차를 설명하는 단일 대표 숫자. 강한 결과는 하나의 숫자가 아니라 반복되는 패턴이다.

주요 한계: 실제 사고가 아니라 저자들의 공격으로 최악 위험을 측정했다. 눈에 띄는 숫자는 설계상 가장 노출된 개인을 묘사한다. 집단별 정확한 격차는 하나의 수치로 요약되기보다 여러 데이터셋에서 일관된 패턴으로 제시된다.

일반 독자는 얼마나 신뢰해도 될까? 전체 평균 프라이버시 지표가 개인 위험을 낮게 평가하고, 남은 위험이 과소대표 환자에게 집중되며, 모델 크기가 커질수록 악화된다는 점에는 높은 신뢰를 둘 수 있다. 현실에서 공격이 얼마나 자주 일어나는지는 이 연구가 측정하지 않았으므로 중간 정도 이하의 확신만 적절하다. 안전한 해석은' 의료 AI 가 당신의 데이터를 흘린다' 도' 프라이버시는 해결됐다' 도 아니다. **표준 프라이버시 검사는 자기 최악의 사례를 보지 못하고, 그 사례는 가장 취약한 환자에게 집중되므로 검사 방식이 바뀌어야 한다.**

출처

Knolle et al., Disparate privacy risks from medical AI, Nature (2026)
https://doi.org/10.1038/s41586-026-10688-0

Technical University of Munich —press release, 'Study reveals privacy risks in medical AI' (2026)
https://www.tum.de/en/news-and-events/all-news/press-releases/details/study-reveals-privacy-risks-in

Medical Xpress (2026) —coverage of the study
https://medicalxpress.com/news/2026-06-patient-groups-vulnerable-privacy-medical.html