
The Clean Paper

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

# A medical AI can be private on average and still expose particular patients — the underrepresented most of all

*A medical-AI model called 'privacy-preserving' usually rests on one average number. This study argues that number is the wrong test. Studying membership inference attacks — which reveal whether a specific person's record was in a model's training data, and so can betray that they had a given disease — the authors measured risk per patient rather than in aggregate, across seven medical datasets (imaging, ECG, electronic records) and many models each. The pattern: models that look safe on average can still let an attacker identify specific individuals almost perfectly (attack AUC ≥ 0.95); the exposed are systematically those from underrepresented groups (minority ethnicity, rare disease, unusual imaging); and it gets worse as models grow. The authors do not say abandon medical AI — they say measure privacy per patient, control model access, and use differential privacy. The uncomfortable core: 'private on average' is not a privacy guarantee, and it fails the patients already least protected.*

---

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

# A privacy guarantee is a promise to a person, not to an average

When a medical-AI model is called "privacy-preserving," that claim usually rests on one number: across all the patients whose data trained the model, the average chance that any individual's membership can be inferred is low. That sounds reassuring. This paper's quiet, uncomfortable point is that it is the wrong number.

Privacy is not an average. It is a promise made to each individual — that being in the training set will not come back to expose them. And an average can keep that promise for almost everyone while breaking it completely for a few. The researchers measured privacy risk one patient at a time, at a resolution no one had used before, and found exactly that: models that look safe in aggregate can leak the membership of specific individuals almost perfectly — and the individuals they leak are, disproportionately, the ones already least protected.

## What the authors did

The team — led by Moritz Knolle, with Daniel Rückert (both at the Technical University of Munich) and Georg Kaissis (Hasso Plattner Institute), alongside colleagues at Imperial College London — took a well-known privacy threat called a **membership inference attack** and changed the question. Instead of asking "on average, how often does this attack succeed across the dataset?", they asked "for *this particular patient*, how exposed are they?"

They ran that per-patient analysis across **seven established medical datasets**, spanning very different data types — medical imaging, electrocardiograms, and electronic health records — and, across them, 200 models each. For every patient in every model, they estimated how confidently an attacker could tell whether that person's record had been part of the training data, then broke the results down by group: disease status, self-reported race, insurance, sex, and imaging protocol.

> **What a membership inference attack actually is**
>
> The leak here is subtler than "the model spits out your record." It is about *membership* — the mere fact that your data was in the training set.
> Start with what one of these models normally does. You hand it a scan — a chest X-ray, say — and it hands back probabilities: a 78% chance of pneumonia, along with readings for cardiomegaly, oedema, consolidation. That everyday diagnostic answer is the *only* thing the attack uses. Nothing exotic.
> The weakness it leans on is this: a model is usually a little **more confident** about the exact examples it was trained on than about ones it has never seen. Think of a student who secretly saw the exam beforehand — on the questions they had practised, the answers come a little too quickly, a little too confidently. Working out, from that giveaway, whether a particular record was one of the examples the model studied is what "membership inference" means.

So here is the attack, step by step. Someone wants to know whether a particular person's scan was used to train the model. They do **not** ask the model for the record — they already hold a candidate scan (the person's own, or a close copy). They send it in once, as an ordinary diagnostic request, and note how confident the answer is. Then they check whether that confidence looks more like a model that *had* trained on the scan or one that *hadn't* — a comparison they can make cheaply by training a stand-in of their own (a "reference model") to learn what that tell-tale over-confidence looks like, on an ordinary computer with no special hardware. If the real model is unusually sure about this scan — as if it recognises it — that is strong evidence the scan was in the training data.

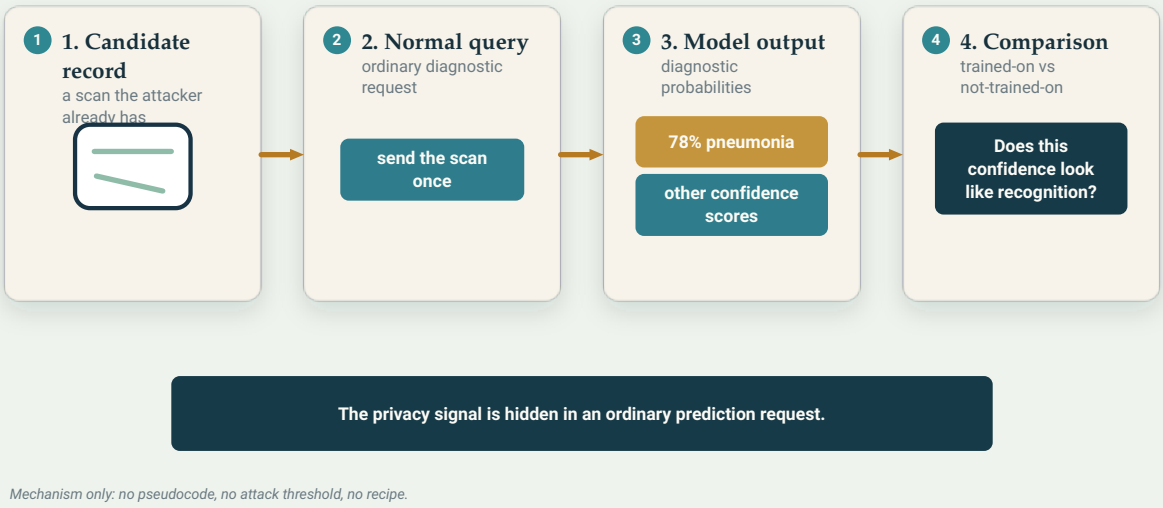
The unsettling part is that nothing about the request looks like an attack: it is the same diagnostic query a clinician would make, and the privacy signal is hidden inside an ordinary prediction. That is exactly why the risk is concrete — even though, so far, it has been demonstrated under stated laboratory assumptions rather than caught in the wild.

Why does membership matter, if the record itself never leaks? Because *membership is a fact*. If a model was trained on patients who received a particular cancer immunotherapy, then confirming that your record is in it reveals that you probably had that cancer — the kind of thing an insurer or employer should never be able to infer. The content stays sealed; the fact of belonging is what escapes.

The authors set out these assumptions plainly — access to the model's ordinary predictions, a candidate record, and the attacker's own reference model — not as a how-to, but so the threat can be reasoned about rather than hand-waved.

## A membership attack can hide inside a normal prediction

*The attacker does not ask for the record. They already hold a candidate and query the model once.*




*Mechanism only: no pseudocode, no attack threshold, no recipe.*

The attack does not need a special privacy API. A normal diagnostic query can leak a membership signal if the model is unusually confident on a record it has seen before.

# What they found

Three findings, each sharper than the last.

**Averages hide the exposed.** Measured in aggregate — the usual way — many of these models look reassuringly private: the attack does no better than chance for most patients. The per-patient view told a different story. As Knolle put it in the study's announcement, previous assessments "have only ever measured the average risk across all patients. We examined the risk at the level of individual patients for the first time — and it paints a very different picture." For some individuals, the attack succeeds **almost perfectly** — the authors measure it as an attack AUC of **0.95 or higher** (0.5 is a coin toss, 1.0 is flawless) — even while the dataset-wide average looked no better than chance.


The average can sit near chance while a small tail of records remains much more exposed. The paper's point is that privacy has to be checked at the patient level, not only in aggregate.



**The exposure is unequal — and it lands on the underrepresented.** The patients most vulnerable to a near-perfect attack were systematically those in groups **underrepresented in the data**: minority ethnicities, rare-disease phenotypes, unusual imaging characteristics. In the electronic-record dataset, Black patients turned up **31% more often** than expected among the most-vulnerable records; in the mammography set, scans flagged as suspicious for malignancy were overrepresented in that danger zone by a striking **+1,179%**. (These two figures are examples, not the whole picture — the paper

reports the same skew across several groups — and they are *relative* over-representations within the small extreme-risk tail: how much more often these patients turn up among the most-exposed records, not the fraction of Black or suspicious-mammography patients who are exposed.) A model has fewer similar examples to blur these patients into, so their records stand out — and standing out is exactly what a membership attack detects. The privacy failure is not random; it concentrates on the people already on the margins.

**Bigger models make it worse.** The number of patients exposed to near-perfect attacks rose sharply with model capacity — in one dermatology dataset, the share climbed from essentially zero in the smallest model to about **one in ten** in the largest. As medical AI models grow larger and more capable — the direction the whole field is moving — this specific risk, the authors warn, gets more severe, not less.

Rückert's summary is blunt: "This is not a tolerable risk. Health data is highly sensitive."

## Why "safe on average" is the wrong test

The temptation is to read a low average risk as a clean bill of health. The paper's core lesson is that averaging here is not merely imprecise — it is measuring the wrong thing.

A privacy guarantee is meaningful only if it holds for the person most at risk, not the person in the middle. A model where 999 in 1,000 patients are unrecoverable but one can be identified almost perfectly is not "99.9% private" in any sense that matters to that one person — and if that one person is predictably the rare-disease patient or the ethnic-minority patient, the metric is not just incomplete, it is quietly discriminatory. It reports safety for the majority and calls it safety for all.

That is the shift the paper forces: from *how private is this model on average?* to *who is the most exposed patient, and who are they?* Those are different questions, and only the second is a privacy question.

## What this does *not* prove — or claim

- It does **not** say medical AI should be abandoned. The authors' framing is mitigation, not retreat: measure and fix the risk, don't stop building.
- It does **not** mean every medical model is leaking, or that any given deployed model has been attacked. It shows the *risk exists and is unequally distributed*, and that standard aggregate metrics miss it.
- It does **not** mean your records are already exposed. The attack needs specific conditions — access to the model, a candidate record, and the attacker's own infrastructure — not a casual capability.
- It does **not** show that patient *content* leaks. What leaks is membership — the fact of inclusion — which is dangerous for a different reason, not because the record is dumped out.
- It does **not** reduce the disparities to a single headline number. The strong, reproducible result is the *pattern* — aggregate metrics understate individual risk, and underrepresented patients bear

the most of it — across datasets and hundreds of models.

## How strong is the evidence?

This is an empirical, methodological result, and a robust one. The pattern — aggregate privacy metrics systematically understate per-patient risk, the residual risk concentrates on underrepresented groups, and it worsens with model capacity — held across **seven datasets of different data types** and a large number of models per dataset. That breadth is exactly what makes a measurement claim credible rather than a one-dataset artefact.

Two honest caveats. First, this is a demonstration of *risk*, measured by running the attacks the authors themselves built; it characterises how well a capable attacker *could* do under stated assumptions, not how often real-world attacks happen. Second, the vivid numbers — near-perfect attack success for some patients — describe the worst-off individuals *by design*; that is the whole point, and they should be read as "the tail is far heavier than the average implies," not as "most patients are exposed."

The appropriate stance is neither alarm nor dismissal: a careful, multi-dataset study showing that the standard way we certify medical-AI privacy is blind to its own worst cases — and that those worst cases fall on the patients with the least margin to spare.

## Why it matters

Two things make this more than a technical footnote.

First, it changes what "privacy-preserving" should be allowed to mean. If a model is released with an average privacy score, that score can be genuinely low and still hide a subset of patients who are near-perfectly identifiable by a membership attack. The paper's practical demand is concrete: assess privacy risk **per patient** before release, control who can access deployed models, and use techniques like **differential privacy** — small, carefully calibrated noise added during training that blunts membership attacks, at a real and managed cost to the model's usefulness (the privacy–utility trade-off is explicit, not free). "We checked the average" should stop counting as having checked.

Second, it braids privacy together with fairness. The same groups that are underrepresented in medical data — and therefore already served worse by medical AI — turn out to be the ones whose privacy that AI protects least. A field working hard on the first inequity cannot treat the second as someone else's department. They are the same people.

The reassuring story of medical-AI privacy — *we measured it, the average is low, we're fine* — is the story this paper takes apart. Not to frighten anyone off the technology, but to move the standard to where it belongs: a promise you can only claim to keep if you have checked it for the person most likely to be hurt.

# Clean summary

Membership inference attacks try to determine whether a specific person's record was in an AI model's training data — and because membership can itself be revealing (that you had a particular disease, say), that is a genuine privacy leak even when the underlying record never surfaces. Prior work measured how often such attacks succeed *on average* across a dataset. This study measured it *per patient*, across seven medical datasets (imaging, ECG, electronic records) and many models each, and found that aggregate metrics badly understate the risk: some individuals can be identified almost perfectly (attack AUC $\geq 0.95$) even when the dataset-wide average looks safe; the most vulnerable are systematically those from underrepresented groups (minority ethnicity, rare disease, unusual imaging); and the problem grows with model size. The authors do not argue for abandoning medical AI — they argue for measuring privacy at the individual level, controlling model access, and using differential privacy. The takeaway: "private on average" is not a privacy guarantee, and the people it fails are usually the ones already least protected.

# No-BS check

**What the paper shows:** Across seven medical datasets and many models each, per-patient membership-inference risk is far higher for some individuals than aggregate metrics suggest — up to near-perfect attack success (AUC $\geq 0.95$) — and that residual risk falls disproportionately on underrepresented groups and grows with model capacity.

**What is plausible but not proven:** That real-world attackers are already exploiting this against deployed clinical models; the study demonstrates the *capability and its distribution* under stated assumptions, not the frequency of attacks in the wild.

**What it does not show:** That medical AI should be abandoned; that every model leaks or that any specific deployed model has been breached; that patient record *contents* (rather than membership) are exposed; a single quantified disparity figure — the robust result is the pattern, not one number.

**Main limitations:** It measures worst-case risk via the authors' own attacks, not observed incidents; the dramatic figures describe the most-exposed individuals by design; and the exact per-group disparities are shown as a consistent pattern across datasets rather than reduced to one number.

**How much confidence should a general reader have?** High that aggregate privacy metrics understate individual risk, that the residual risk concentrates on underrepresented patients, and that this worsens with model size — it is demonstrated across many datasets and models. Moderate on the real-world frequency of such attacks, which this study does not measure. The safe reading: not "medical AI leaks your data," and not "privacy is solved," but "the standard privacy check is blind to its own worst cases, and those cases fall on the most vulnerable patients — so the check has to change."

## Sources

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