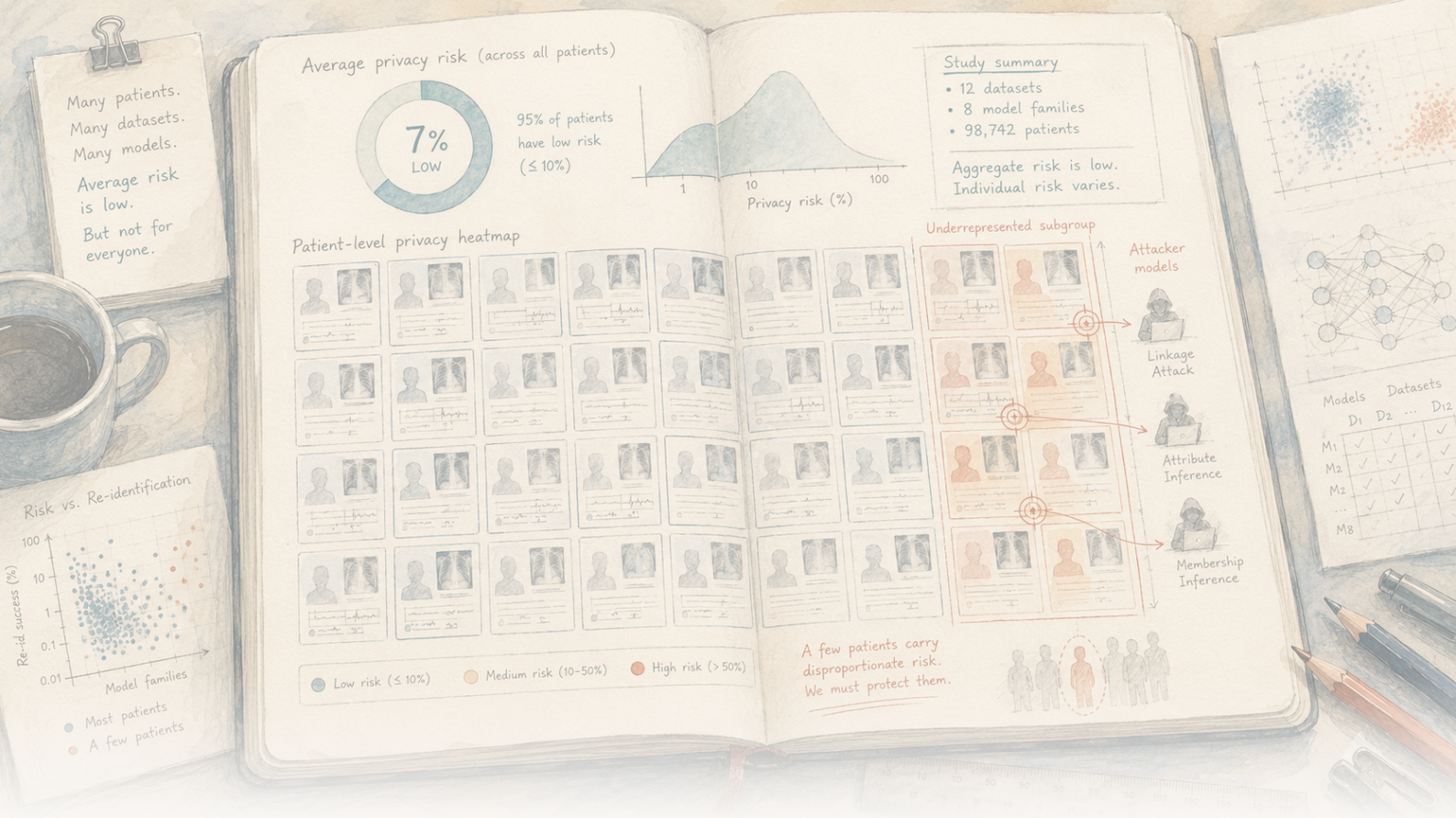



The Clean Paper

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

# Medical AI lè jẹ́ private on average, ṣùgbọ́n ṣì expose particular patients — underrepresented most of all

*Medical-AI model tí a pe 'privacy-preserving' sábà rest lórí average number kan. Study yìí argue number yẹn wrong test. Studying membership inference attacks — which reveal whether a specific person's record was in training data and can betray sensitive disease membership — authors measured risk per patient rather than aggregate across seven medical datasets and many models. Pattern: models safe on average can let attacker identify specific individuals almost perfectly (AUC ≥ 0.95); exposed are systematically underrepresented groups; risk worsens as models grow. Authors don't say abandon medical AI — measure privacy per patient, control access, use differential privacy. Core: 'private on average' is not privacy guarantee.*

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

# Privacy guarantee jẹ́ ìlérí sí ènìyàn kan, kì í ṣe sí average

Nígbà tí a bá pe medical-AI àwòṣe kan ní “privacy-preserving,” claim náà sábà rest lórí number kan: across gbogbo aláìsàn tí dátà wọn train àwòṣe, average chance pé membership individual kan lè be inferred low. Èyí sounds reassuring. Quiet, uncomfortable kókó ìwé ìwádìí yìí ni pé wrong number ni.

Privacy kì í ṣe average. Ó jẹ́ promise sí individual kọ̀ọ̀kan — pé being in training set kò ní come back to expose wọn. Average lè keep promise fún almost everyone, ṣùgbọ́n break completely fún few. olùwádìí measure àṣírí ewu ọ̀kan aláìsàn at a àkókò, at resolution no ọ̀kan lò ṣáájú, wọ́n sì rí exactly that: àwòṣe tí look láìléwu in aggregate lè leak membership of specific individuals almost perfectly — individuals tí wọ́n leak sì disproportionately jẹ́ those already least protected.

# Ohun tí àwọn olùkọ̀wé ṣe

ẹgbẹ́ — led by Moritz Knolle, pẹ̀lú Daniel Rückert (Technical University of Munich) àti Georg Kaissis (Hasso Plattner Institute), plus colleagues at Imperial College London — gba well-known àṣírí threat kan, **membership inference ìkọlù**, wọ́n sì yí ìbéèrè padà. Dípò “on average, how often does ìkọlù succeed across dataset?”, wọ́n béèrè “fún *aláìsàn yìí pàtó*, how exposed are they?”

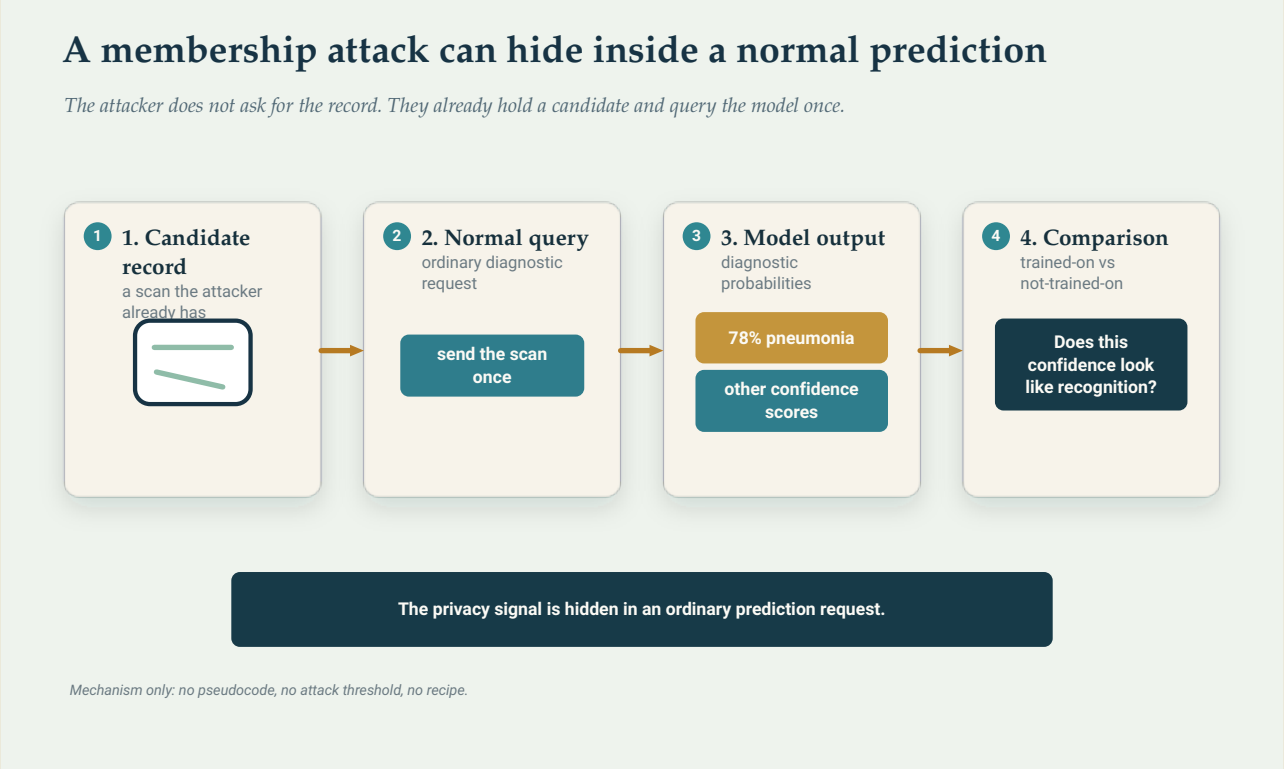
Wọ́n run per-patient ìtúpalẹ̀ kọjá **seven established ìṣègùn datasets**, dátà types very yàtọ̀ — ìṣègùn imaging, electrocardiograms, electronic health àkọsílẹ̀ — àti 200 àwòṣe each. Fún aláìsàn kọ̀ọ̀kan nínú àwòṣe kọ̀ọ̀kan, wọ́n ìṣírò àfojúsùn how confidently attacker lè tell whether àkọsílẹ̀ person yẹn part of training dátà, then break èsì down by ẹgbẹ́: disease status, self-reported race, insurance, sex, imaging protocol.

> **Kí ni membership inference attack gidi?**
>
> Leak níbí subtle ju “àwòṣe spits out your àkọsílẹ̀.” Ó jẹ́ nípa *membership* — fact pé dátà rẹ wà nínú training set.
> Bẹ̀rẹ̀ pẹ̀lú ohun tí àwòṣe báyìí normally ṣe. O fún un ní scan — chest X-ray fún example — ó sì give probabilities: 78% chance pneumonia, plus readings fún cardiomegaly, oedema, consolidation. Everyday diagnostic answer yẹn nìkan ni ìkọlù lo. Nothing exotic.
> Weakness: àwòṣe sábà **sí i confident** díẹ̀ lórí exact examples tí a train rẹ̀ on ju ones never seen. Think of student tí secretly saw exam ṣáájú — on practised ìbéèrè answers come too quickly, too confidently. Figuring out láti that giveaway whether particular àkọsílẹ̀ was ọ̀kan of examples àwòṣe studied ní “membership inference.”
> Attack step by step: attacker wants to know whether particular person’s scan lò to train àwòṣe. Wọn **kò** ask àwòṣe fún àkọsílẹ̀ — wọn already hold candidate scan (person own tàbí close copy). Wọn send it once as ordinary diagnostic request, note ìgbẹ́kẹ̀lé. Then fi wé whether ìgbẹ́kẹ̀lé looks sí i like àwòṣe that *had* trained on scan tàbí ọ̀kan that *hadn’t* — ìfiwéra they lè make cheaply by training own stand-in (“reference àwòṣe”) to learn tell-tale over-confidence on ordinary computer. If gidi àwòṣe unusually sure nípa scan — as if recognizes it — lágbára ẹ̀rí scan was in training dátà.
> Unsettling part: request looks exactly like normal clinician query; àṣírí àmì hidden inside ordinary

prediction. That makes ewu concrete — though demonstrated lábẹ́ stated lab assumptions, not caught in wild.

Why membership matters if àkọsílẹ̀ itself never leaks? Because *membership is a fact*. If àwòṣe trained on aláìsàn receiving particular cancer immunotherapy, confirming your àkọsílẹ̀ is in it reveals you probably had that cancer — information insurer/employer yẹ kí never infer. Content stays sealed; fact of belonging escapes.

òǹkọ̀wé ipò assumptions plainly — access to ordinary àwòṣe predictions, candidate àkọsílẹ̀, attacker's own reference àwòṣe — not as how-to, ṣùgbọ́n so threat lè be reasoned nípa rather than hand-waved.




Attack kò need special àṣírí API. Normal diagnostic query lè leak membership àmì if àwòṣe unusually confident on àkọsílẹ̀ it has seen ṣáájú.



# Ohun tí wọ́n rí

Findings mẹ́ta, each sharper than last.

**Averages hide exposed aláìsàn.** Measured in aggregate — usual way — many àwòṣe look reassuringly private: ìkọlù no better than chance fún most aláìsàn. Per-patient view yàtọ̀. As Knolle put it in ìwádìí announcement, previous assessments "have nìkan ever measured the average ewu across gbogbo aláìsàn. We examined the ewu at the ìpele of individual aláìsàn fún the first àkókò — àti it paints a very yàtọ̀ picture." Fún some individuals, ìkọlù

succeeds **almost perfectly** — ìkọlù AUC **0.95 tàbí higher** (0.5 coin toss, 1.0 flawless) — even while dataset-wide average looked no better than chance.


Average lè sit near chance while small tail of àkọsílẹ̀ remains much sí i exposed. ìwé ìwádìí kókó: àṣírí has to be checked at aláìsàn ìpele, not nìkan aggregate.

Original Aurora diagram — The Clean Paper CC BY 4.0

**Exposure unequal — lands on underrepresented.** aláìsàn most vulnerable to near-perfect ìkọlù were systematically láti ẹgbẹ́ **underrepresented in dátà**: minority ethnicities, rare-disease phenotypes, unusual imaging characteristics. Nínú electronic-record dataset, Black aláìsàn appeared **31% sí i often** than expected among most-vulnerable àkọsílẹ̀; nínú mammography set, scans suspicious fún malignancy overrepresented in danger zone by **+1,179%**. (Figures are examples, not whole picture; kan náà skew across several ẹgbẹ́; they are *relative* over-representations nínú small extreme-risk tail — how much sí i often these aláìsàn appear among most-exposed àkọsílẹ̀, not fraction of Black tàbí suspicious-mammography aláìsàn exposed.) àwòṣe has fewer similar examples to blur these aláìsàn sínú, so àkọsílẹ̀ stand out — exactly what membership ìkọlù detects. Privacy ìkùnà not random; it concentrates on people already at margins.

**Bigger àwòṣe make it worse.** Number aláìsàn exposed to near-perfect ìkọlù rose sharply pẹ̀lú àwòṣe capacity — nínú dermatology dataset, share climbed láti essentially zero in smallest àwòṣe to nípa **ọ̀kan in ten** in largest. As ìṣègùn AI àwòṣe get larger/sí i capable, òǹkọ̀wé warn this ewu gets sí i severe, not less.

Rückert summary blunt: “èyí kì í ṣe a tolerable ewu. Health dátà is highly sensitive.”

# Kí ló dé tí “láìléwu on average” fi wrong ìdánwò?

Temptation ni lati read low average ewu as clean bill of health. Core lesson ìwé ìwádìí: averaging here not merely imprecise — it measures wrong thing.

Privacy guarantee meaningful nìkan if holds fún person most at ewu, not person in middle. àwòṣe where 999 in 1,000 aláìsàn unrecoverable ṣùgbọ́n ọ̀kan identified almost perfectly kì í “99.9% private” in sense that matters to that ọ̀kan person — àti if that ọ̀kan person predictably rare-disease aláìsàn tàbí ethnic-minority aláìsàn, metric not just incomplete, quietly discriminatory. It reports àábò fún majority àti calls it àábò fún gbogbo.

Shift ìwé ìwádìí forces: láti *how private is àwòṣe on average?* to *who is most exposed aláìsàn, àti who are they?* Different ìbéèrè, nìkan second is àṣírí ìbéèrè.

# Ohun tí èyí kò fi ẹ̀rí múlẹ̀ — tàbí claim

- Kò say ìṣègùn AI yẹ kí be abandoned. òǹkọ̀wé framing is mitigation, not retreat: measure àti fix ewu, don’t stop building.
- Kò mean every ìṣègùn àwòṣe leaking, tàbí any deployed àwòṣe attacked. It fi hàn *ewu exists àti unequally distributed*, standard aggregate metrics miss it.
- Kò mean your àkọsílẹ̀ already exposed. Attack needs specific conditions — àwòṣe access, candidate àkọsílẹ̀, attacker infrastructure — not casual capability.
- Kò fi hàn aláìsàn *content* leaks. Membership — fact of inclusion — leaks, dangerous fún yàtọ̀ reason, not àkọsílẹ̀ dump.
- Kò dín kù disparities to ọ̀kan àkọlé number. Strong reproducible èsì is *àpẹẹrẹ* — aggregate metrics understate individual ewu, underrepresented aláìsàn bear most — across datasets àti hundreds àwòṣe.

# Báwo ni ẹ̀rí ṣe lágbára tó?

Empirical methodological èsì, robust. Pattern — aggregate àṣírí metrics understate per-patient ewu, residual ewu concentrates on underrepresented ẹgbẹ́, worsens pẹ̀lú àwòṣe capacity — held across **seven datasets of yàtọ̀ dátà types** àti many àwòṣe. Breadth is what makes measurement claim credible rather than one-dataset artefact.

Two honest caveats. First, demonstration of *ewu*, measured by ìkọlù òǹkọ̀wé built; it characterises how well capable attacker *lè* do lábẹ́ assumptions, not ayé gidi ìkọlù frequency. Second, vivid numbers — near-perfect success fún some aláìsàn — describe worst-off individuals *by àpẹẹrẹ*; whole kókó, yẹ kí be read as “tail much heavier than average implies,” not “most aláìsàn exposed.”

Appropriate stance neither alarm nor dismissal: careful multi-dataset ìwádìí showing standard way we certify medical-AI àṣírí blind to worst ọ̀ràn — àti worst ọ̀ràn fall on aláìsàn pẹ̀lú least margin to spare.

# Kí nìdí tí ó fi ṣe pàtàkì?

Two things make this sí i than technical footnote.

First, it changes what “privacy-preserving” yẹ kí be allowed to mean. àwòṣe lè have genuinely low average àṣírí score àti still hide subset of aláìsàn near-perfectly identifiable by membership ìkọlù. Practical demand concrete: assess àṣírí ewu **per aláìsàn** ṣáájú release, control access to deployed àwòṣe, lò techniques like **differential àṣírí** — small carefully calibrated ariwo added during training to blunt membership ìkọlù, at gidi managed cost to àwòṣe usefulness (àṣírí–utility trade-off explicit, not free). “We checked average” yẹ kí stop counting as checked.

Second, it braids àṣírí pẹ̀lú fairness. Same ẹgbẹ́ underrepresented in ìṣègùn dátà — already served worse by ìṣègùn AI — turn out to have least àṣírí protection. Field working hard on first inequity kò lè treat second as someone else’s department. Same people.

Reassuring ìtàn — *we measured it, average low, we’re fine* — is ìtàn ìwé ìwádìí dismantles. Not to scare people láti technology, ṣùgbọ́n to move standard where it belongs: a promise you lè claim nìkan if checked fún person most likely harmed.

# Àkótán kedere

Membership inference ìkọlù try determine whether specific person’s àkọsílẹ̀ was in AI training dátà — membership itself lè reveal sensitive facts even when àkọsílẹ̀ never ojú. Prior iṣẹ́ measured ìkọlù success *on average*. ìwádìí yìí measured *per aláìsàn* across seven ìṣègùn datasets (imaging, ECG, electronic àkọsílẹ̀) àti many àwòṣe, rí aggregate metrics badly understate ewu: some individuals identified almost perfectly (ìkọlù AUC $\geq$ 0.95) even when dataset average láìléwu; most vulnerable systematically láti underrepresented ẹgbẹ́ (minority ethnicity, rare disease, unusual imaging); ìṣòro grows pẹ̀lú àwòṣe ìwọ̀n. òǹkọ̀wé do not argue abandon ìṣègùn AI — they say measure àṣírí per aláìsàn, control àwòṣe access, lò differential àṣírí. Takeaway: “private on average” kì í ṣe àṣírí guarantee, àti ìkùnà hit people already least protected.

# Àyẹ̀wò láìsí àṣejù

**Ohun tí ìwé ìwádìí fi hàn:** Across seven ìṣègùn datasets àti many àwòṣe, per-patient membership-inference ewu much higher fún some individuals than aggregate metrics tọ́ka sí — up to near-perfect ìkọlù success (AUC $\geq$ 0.95) — residual ewu disproportionately falls on underrepresented ẹgbẹ́ àti grows pẹ̀lú àwòṣe capacity.

**Ohun tó ṣeé gbà ṣùgbọ́n tí a kò fi ẹ̀rí múlẹ̀:** That ayé gidi attackers already exploit this against deployed ìṣègùn àwòṣe; ìwádìí demonstrates *capability àti distribution* lábẹ́ assumptions, not frequency in wild.

**Ohun tí kò fi hàn:** Medical AI yẹ kí be abandoned; every àwòṣe leaks tàbí specific deployed àwòṣe breached; aláìsàn àkọsílẹ̀ *contents* rather than membership exposed; single quantified disparity figure — robust èsì is àpẹẹrẹ, not ọ̀kan number.

**Main limitations:** Measures worst-case ewu via òǹkọ̀wé’ ìkọlù, not observed incidents; dramatic figures describe

most-exposed individuals by àpẹrẹ; exact per-group disparities shown as consistent àpẹẹrẹ across datasets, not ọ̀kan number.

**Confidence wo ni gbogbogbò reader yẹ kí ó ní?** High that aggregate àṣírí metrics understate individual ewu, residual ewu concentrates on underrepresented aláìsàn, worsens pẹ̀lú àwòṣe ìwọ̀n — demonstrated across many datasets/àwòṣe. Moderate on ayé gidi frequency of ìkọlù, not measured. Safe reading: not “ìṣègùn AI leaks your dátà,” not “àṣírí solved,” ṣùgbọ́n “standard àṣírí check is blind to own worst ọ̀ràn, àti those ọ̀ràn fall on most vulnerable aláìsàn — so check gbọ́dọ̀ change.”

---

## Àwọn orísun


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