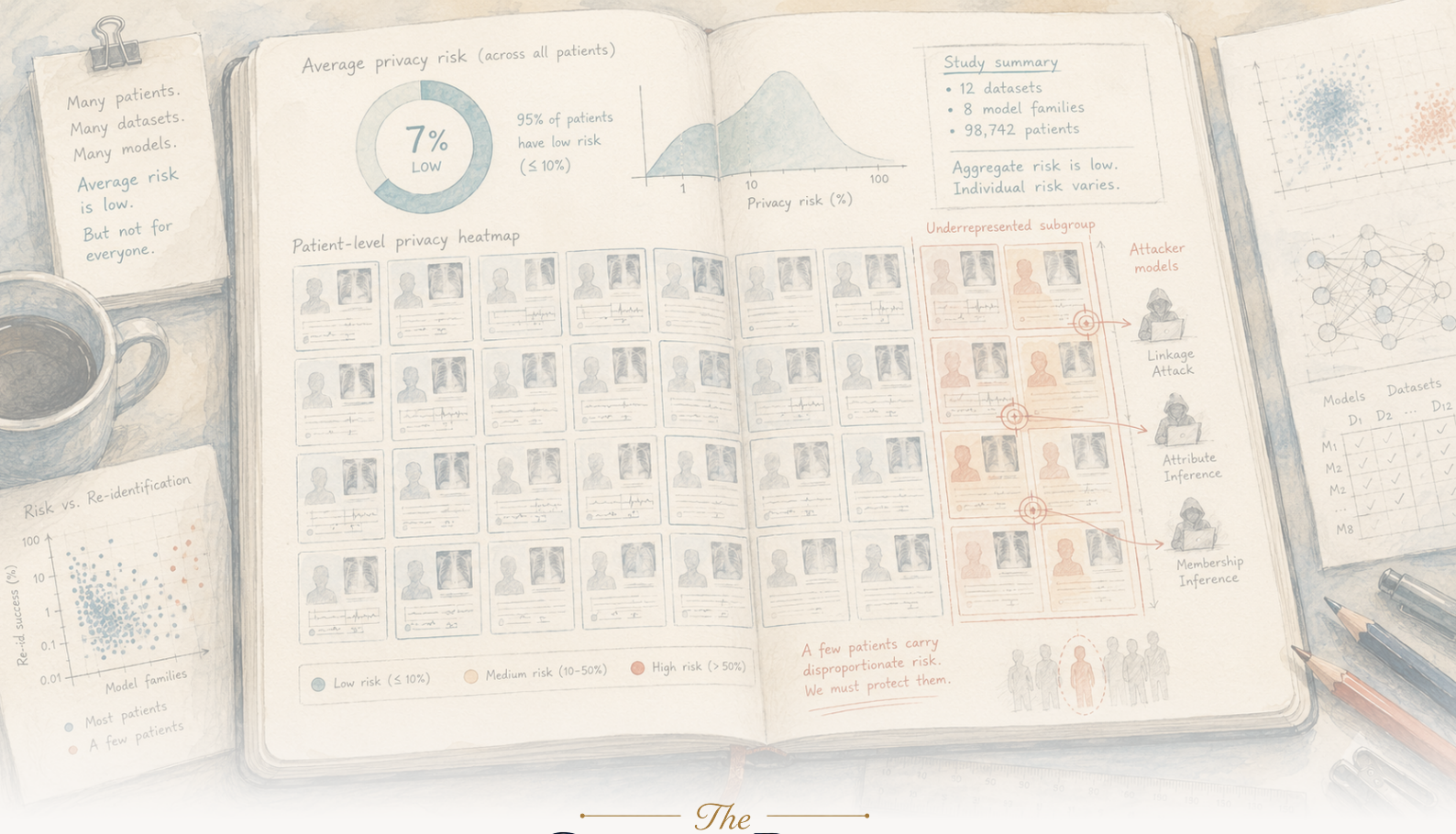



The Clean Paper

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

# Medical AI average irratti private ta'ee illee patients specific expose gochuu danda'a — keessumaa underrepresented

*Medical-AI model 'privacy-preserving' jedhu yeroo baay'ee average number tokko irratti hirkata. Study kun wrong test jedha. Membership inference attacks — record person specific training data keessa ture reveal godhu — per patient safaruun, seven medical datasets (imaging, ECG, EHR) fi models hedduu keessatti, average safe fakkaatu illee individuals tokko tokko almost perfectly identify ta'uu (AUC ≥0.95), exposed systematic underrepresented groups, problem model growth waliin worse ta'uu argan. Authors abandon medical AI hin jedhu — privacy per patient measure, access control, differential privacy jedhu. Core: 'private on average' privacy guarantee miti, already least-protected patients irratti fail godha.*

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

# iccitii dhuunfaa guarantee giddugaleessa'f miti — nama tokkoon tokkoof waadaa dha

Medical-AI moodeela "iccitii dhuunfaa-preserving" jedhamee yeroo waamamu himannaa kun yeroo baay'ee lakkoofsa tokko irratti hirkata: dhukkubsattoota daataa isaanii moodeela leenjise hunda gidduutti, giddugaleessa chance miseensummaa nama dhuunfaa tilmaamuu xiqqaa dha. Nama tasgabbeessu fakkaata. Barruu kanaa qabxii tasgabbaa'aa, uncomfortable: lakkoofsa dogoggoraa dha.

iccitii dhuunfaa giddugaleessa miti. Nama tokkoon tokkoof waadaa dha — leenjii set keessa turuun isaanii gara fuulduraatti akka isaan expose hin goone. giddugaleessa waadaa sana jechuun ni danda'ama everyone'f eeguu danda'a, muraasaaf guutumaan cabsuu illee. Researchers patient-by-patient iccitii dhuunfaa balaa resolution kanaan dura hin fayyadamne irratti safaranii exactly kana argan: walitti qabame irratti nageenya qabu fakkaatan moodeelota addaa individuals mis-eensummaa jechuun ni danda'ama perfectly leak gochuu danda'u — saaxilame individuals immoo disproportionally duraan least protected gareewwan irraa.

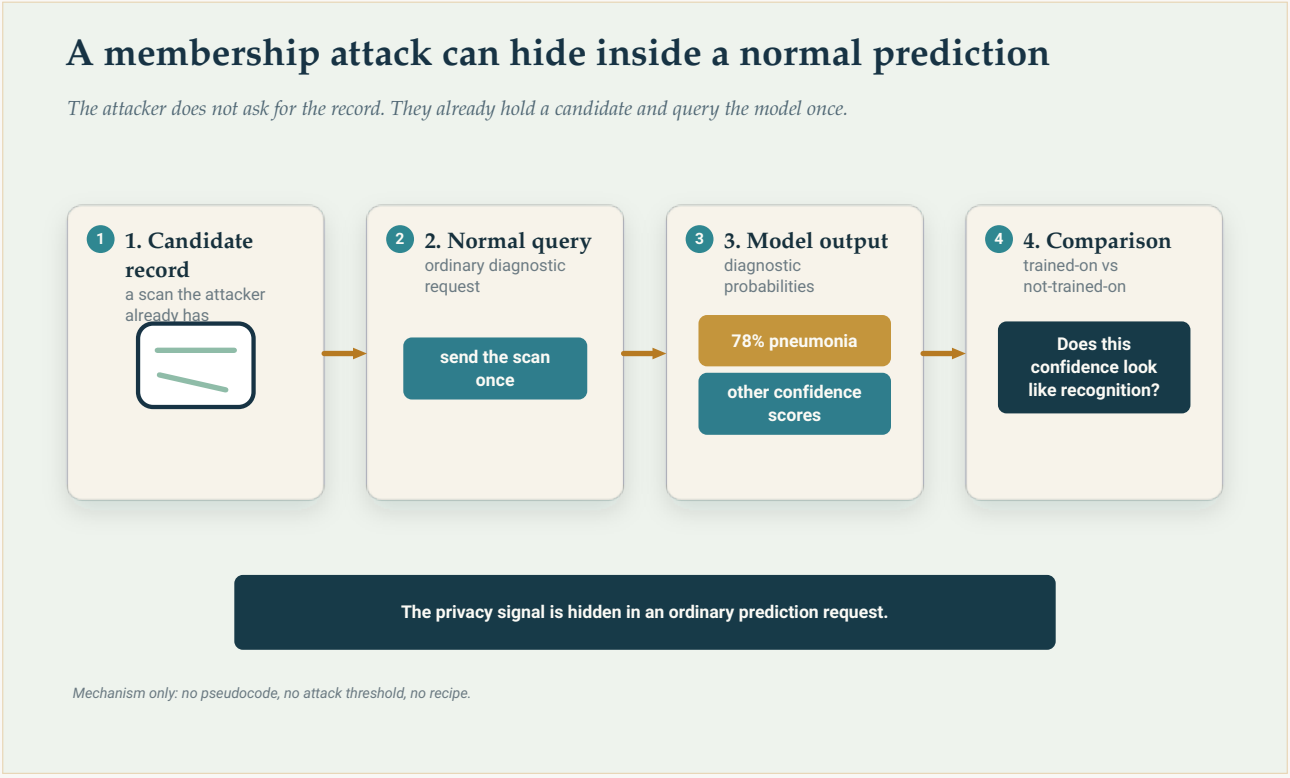
## Barreessitoonni maal godhan

Gareen Moritz Knolle led, Daniel Rückert, Georg Kaissis fi colleagues miseensummaa inference haleellaa beekamaa tokko fudhatee gaaffii jijjiire. "giddugaleessa keessatti haleellaa kuusaa daataa irratti hammam milkaa'a?" osoo hin taane "*dhukkubsataa kana particular* hammam saaxilame?" gaafatan.

dhukkubsataa tokkoon tokkoon xiinxala **yaalaa kuusaa daataa established torba** — imaging, electrocardiograms, electronic fayyaa galmeewwan — irratti, tokkoon tokkoon kuusaa daataa keessatti moodeelota 200, gaggeessan. Dhukkubsataa hunda moodeela hunda keessatti haleellaa raawwatu galmee isaanii leenjii daataa keessaa ta'uu amanamummaa hammam qabaachuu akka danda'u tilmaama; bu'aawwan garee'n qoodan: dhibee status, self-reported race, insurance, sex, imaging protocol.

> **Maal a miseensummaa inference haleellaa actually is**
>
> Leak as jiru "moodeela galmee kee baasa" caalaa subtle. Inni *miseensummaa* — daataa kee leenjii set keessa ture fact — irratti.
> Moodeela adda baasuu dhibee normal maal godha? Scan chest X-ray kennita, carraawwan deebisa: pneumonia 78%, cardiomegaly, oedema, consolidation readings. haleellaa exactly normal bu'aa kana qofa fayyadama.
> Weakness: moodeela fakkeenyonni sirrii leenjii irratti unseen fakkeenyonni caalaa yeroo baay'ee xiqqoo **confident caalaa** ta'a. Student exam dura gaaffilee arge fakkaata — practised gaaffilee irratti deebii too quickly/too confidently. Galmee particular tokko fakkeenyonni moodeela studied keessaa ture moo amanamummaa giveaway irraa infer gochuun miseen-

summaa inference dha.
haleellaa conceptual: haleellaa raawwatu kan filatame scan qaba, moodeela irraa galmee gaafatu miti. Idilee gaaffii adda baasuu dhibee godhee amanamummaa note godha. Sana booda amanamummaa sun moodeela scan sana irratti *leenjifame* ta'e vs *hin leenjifame* moodeela amala kam fakkaata wal bira qaba — mataa isaa wabii moodeela train godhee akkaataa over-amanamummaa barachuudhaan. Dhugaa moodeela unusually sure yoo ta'e, leenjii miseensummaa ragaa cimaa.
Unsettling kutaa request haleellaa fakkaachuu hin qabu: wal-fakkaataa clinician query, iccitii dhuunfaa mallattoo idilee tilmaama keessa dhokata. Balaa ifa ta'e dha, vaikka lab yaadota bu'uuraa jalatti demonstrated, wild incidents miti.
miseensummaa maaliif barbaachisaa, galmee qabiyyee leak hin taane? miseensummaa ofii fact dha. Moodeela dhukkubsattoota cancer immunotherapy argatan irratti leenjifame yoo ta'e, galmee kee leenjii keessa jira confirm gochuun cancer sana qabaachuu malte mul'isa — insurer/employer akka hin infer goone. Qabiyyee sealed, belonging fact escapes.
Barreessitoonni yaadota bu'uuraa ifatti kaa'u — moodeela idilee tilmaamota access, kan filatame galmee, haleellaa raawwatu's wabii moodeela — threat sababa gochuuf, how-to exploit kennuuf miti.




haleellaa addaa iccitii dhuunfaa API hin barbaadu. Idilee gaaffii adda baasuu dhibee moodeela galmee duraan arge irratti unusually confident yoo ta'e miseensummaa mallattoo leak gochuu danda'a.

Original Aurora diagram — The Clean Paper CC BY 4.0

# Maal argatan

Argannoowwan sadii, tokko tokko caalaa sharp.

**Averages saaxilame people dhoksu.** Qorannoo announcement keessatti illee akka ibsametti, walitti qabame — usual safartuu — keessatti moodeelota hedduun nama tasgabbeessu dhuunfaa fakkaatu: haleellaa dhukkubsattoota baay'ee irratti chance caalaa hin caalu. dhukkubsataa tokkoon tokkoon view adda. Individuals tokko tokko irratti haleellaa **jechuun ni danda'ama perfectly** milkaa'a — haleellaa AUC **0.95 yookaan ol** (0.5 coin toss, 1.0 flawless) — even kuusaa daataa giddugaleessa chance-level fakkaatu.




giddugaleessa chance naannoo taa'uu danda'a, garuu galmeewwan muraasa gara dhumaa baay'ee saaxilame ta'ee hafa. Barruu qabxii: iccitii dhuunfaa dhukkubsataa sadarkaa irratti qorachuu qaba, walitti qabame qofa miti.

Original Aurora diagram — The Clean Paper CC BY 4.0

**Saaxilamuu unequal — bakki isaanii xiqqaa ta'e irratti concentrate.** sirrummaa guutuutti dhihoo haleellaa vulnerable dhukkubsattoota systematic gareewwan **daataa keessatti bakki isaanii xiqqaa ta'e**: minority ethnicities, rare-disease phenotypes, unusual imaging characteristics. Electronic-record kuusaa daataa keessatti Gurraacha dhukkubsattoota most-vulnerable galmeewwan keessa eegamu caalaa **31% caalaa often**; mammography set keessatti scans suspicious for malignancy danger gara dhumaa keessatti **+1,179%** overrepresented. Fakkiiwwan kun fakkeenyonni qofa fi *walbira qabamee over-representations keessatti xiqqaa extreme-risk gara dhumaa*: saaxilame dhukkubsattoota share miti. Moodeela similar fakkeenyonni xiqqaa qaba, galmeewwan isaanii standout; miseensummaa haleellaa exactly standout adda baasa. iccitii dhuunfaa kufaatii tasaa miti; people margins irratti concentrate.

**Moodeelota guddaan caalaatti hamaa godhu.** sirrummaa guutuutti dhihoo attacks saaxilame dhukkubsattoota moodeela dandeettii waliin sharply dabalan — dermatology kuusaa daataa tokko keessatti isa xiqqaa moodeela essentially zero irraa isa guddaa keessatti **~tokko in ten**. Damee guddaa caalaa/capable moodeelota garaatti deemu waliin balaa addaa kun caalaatti hamaa ta'uu barreessitoonni warn.

Rückert akkana jechuun cuunfe: "Balaan kun fudhatama hin qabu. Daataan fayyaa baay'ee miira namaa tuqa."

## Maaliif "nageenya qabu on giddugaleessa" qormaata dogoggoraa dha

Gadi-aanaa giddugaleessa balaa ragaa rakkoon hin jirre akka ta'etti dubbisuun temptation dha. Core lesson: averaging imprecise qofa miti — wrong gaaffii safara.

iccitii dhuunfaa waadaan hiika qabu ta'uuf nama balaa guddaa keessa jiru irratti hojii irra ooluu qaba. Moodeela dhukkubsattoota 1,000 keessaa 999 unrecoverable garuu tokko jechuun ni danda'ama perfectly identified ta'u "99.9% dhuunfaa" jechuun nama sanaaf meaning hin qabu. Namni tokko predictable rare-disease/ethnic-minority dhukkubsataa yoo ta'e safartuu incomplete qofa miti, quietly discriminatory dha. Majority nageenya gabaasa godhee hunda nageenya jedha.

Shift barruu: *moodeela giddugaleessa irratti hammam dhuunfaa?* irraa *dhukkubsataa irra caalaa saaxilame eenyu, eenyu irraa?* gaaffiitti. Lamaan adda; iccitii dhuunfaa gaaffii dhugaa lammaffaa dha.

## Kun maal *hin* mirkaneessu — yookaan himannaa hin godhu

- Yaalaa AI abandon godhaa hin jedhu. Framing mitigation dha: balaa safaruu/fooyyessuu, building hin dhaabin.
- Every moodeela leaking yookaan deployed moodeela haleellaa ta'e hin jedhu. Balaa exists/unequally distributed fi walitti qabame safartuuwwan miss godhu agarsiisa.
- Galmeewwan kee duraan saaxilame jechuu miti. haleellaa haalawwan addaa — moodeela access, kan filatame galmee, haleellaa raawwatu's infrastructure — barbaada.
- Dhukkubsataa *qabiyyee* leak hin agarsiisu. miseensummaa fact inclusion leak dha; adda danger.
- Disparities tokko mata-duree numbertti reduce hin godhu. jabaataa bu'aa *akkaataa*: walitti qabame nama dhuunfaa balaa understate, bakki isaanii xiqqaa ta'e irra caalaa burden, kuusaa daataa/moodeelota keessatti.

# Ragaan hammam cimaa dha?

Ragaa irratti hundaa'e methodological bu'aa jabaataa dha. Akkaataa — walitti qabame safartuuwwan dhukkubsataa tokkoon tokkoon balaa understate, residual balaa bakki isaanii xiqqaa ta'e gareewwan irratti, moodeela dandeettii waliin caalaatti hamaa — **kuusaa daataa torba gosoota daataa adda** fi moodeelota hedduu keessatti itti fufe. Breadth safartuu himannaa credible godha.

Caveats lama. Tokkoffaa balaa demonstration, barreessitoonni built attacks raawwachuun; capable haleellaa raawwatu yaadota bu'uuraa jalatti *could* do maal safara, haala addunyaa dhugaa frequency miti. Lammaffaa sirrummaa guutuutti dhihoo vivid lakkoofsota worst-off individuals *akkaataa qophii isaatiin* describe; qabxii "gara dhumaa giddugaleessa caalaa heavy" dha, "irra caalaa dhukkubsattoota saaxilame" miti.

Stance alarm/dismissal miti: multi-kuusaa daataa qorannoo istaandaardii medical-AI iccitii dhuunfaa certification isa hamaa haalawwan ofii hin argine, isa hamaa haalawwan least margin people irratti fall.

# Maaliif barbaachisaa dha

Wantoonni lama teknikaa footnote caalaa godhu.

Tokkoffaa, "iccitii dhuunfaa-preserving" maal jechuu allowed akka ta'u jijjiira. giddugaleessa iccitii dhuunfaa score gadi-aanaa ta'ee illee sirrummaa guutuutti dhihoo miseensummaa identifiable subset dhoksuu danda'a. Hojiirra oolu demand: gadhiisuu dura iccitii dhuunfaa balaa **tokkoon tokkoon dhukkubsataa** assess, deployed moodeela access to'annoo, techniques akka **differential iccitii dhuunfaa** — leenjii keessatti calibrated noise miseensummaa attacks blunt godhu, dhugaa iccitii dhuunfaa–utility trade-off waliin — fayyadami. "giddugaleessa sakattaane" checking ta'uu dhiisuu qaba.

Lammaffaa, iccitii dhuunfaa fi fairness walitti hida. Gareewwan yaalaa daataa keessatti bakki isaanii xiqqaa ta'e — duraan AI'n caalaatti hamaa served — iccitii dhuunfaa isaanii illee least protected. Damee jalqabaa inequity irratti hojjetu lammaffaa someone else's department jechuun hin danda'u. Wal-fakkaataa people dha.

Nama tasgabbeessu seenaa "giddugaleessa gadi-aanaa, fine" barruu kun diiga. Technology irraa nama sodaachisuuf miti, istaandaardii person irra caalaa carraan isaa guddaa harmed irratti qoratamee qofa iccitii dhuunfaa promise jedhu garaatti sochoosuuf.

# Cuunfaa gabaabaa

miseensummaa inference attacks galmee addaa tokko AI leenjii daataa keessa ture moo murteessu; miseensummaa ofii dhibee seenaa mul'isuu danda'a, galmee qabiyyee hin baane illee. Duraa

hojii haleellaa milkaa'ina *giddugaleessa* kuusaa daataa irratti safara. Qorannoo kun *tokkoon tokkoon dhukkubsataa*, yaalaa kuusaa daataa torba (imaging, ECG, EHR) fi moodeelota hedduu keessatti safare. Walitti qabame safartuuwwan balaa baay'ee understate: individuals tokko tokko jechuun ni danda'ama perfectly identify ta'uu danda'u (haleellaa AUC $\geq$ 0.95) vaikka kuusaa daataa giddugaleessa nageenya qabu; vulnerable systematic bakki isaanii xiqqaa ta'e gareewwan (minority ethnicity, muraasa dhibee, unusual imaging); rakkoo moodeela hammamtaa waliin caalaatti hamaa. Barreessitoonni yaalaa AI abandon hin jedhu — iccitii dhuunfaa tokkoon tokkoon dhukkubsataa measure, moodeela access to'annoo, differential iccitii dhuunfaa fayyadama jedhu. ergaa ijoo: "dhuunfaa on giddugaleessa" iccitii dhuunfaa guarantee miti, fail gochuun duraan least protected dhukkubsattoota irratti.

# Sakatta'a ifaa

**barruun maal agarsiisa:** Seven yaalaa kuusaa daataa/moodeelota hedduu keessatti dhukkubsataa tokkoon tokkoon miseensummaa-inference balaa individuals tokko tokkoof walitti qabame safartuuwwan caalaa baay'ee ol, sirrummaa guutuutti dhihoo haleellaa AUC $\geq$0.95; residual balaa disproportionately bakki isaanii xiqqaa ta'e gareewwan irratti, moodeela dandeettii waliin grows.

**Waan amansiisaa ta'e garuu hin mirkanoofne:** haala addunyaa dhugaa attackers duraan deployed kilinikaa moodeelota irratti exploit gochaa jiraachuu; qorannoo dandeettii/distribution yaadota bu'uuraa jalatti, haleellaa frequency wild keessatti miti.

**Waan hin agarsiifne:** Yaalaa AI abandon; every moodeela leaks/deployed breach; dhukkubsataa galmee *contents* saaxilame; tokko universal disparity lakkoofsa.

**Daangaawwan ijoo:** Barreessitoonni mataa isaa attacks'n haala hamaa keessaa isa hamaa balaa safara, mul'ate incidents miti; dramatic lakkoofsota most-exposed akkaataa qophii isaatiin; sirrii garee disparities akkaataa wal-simu, tokko lakkoofsa miti.

**Dubbisaan waliigalaa amanamummaa hammam qabaachuu qaba?** Walitti qabame iccitii dhuunfaa safartuuwwan nama dhuunfaa balaa understate, residual balaa bakki isaanii xiqqaa ta'e dhukkubsattoota irratti concentrate, moodeela hammamtaa waliin caalaatti hamaa ta'uu irratti ol-aanaa. haala addunyaa dhugaa haleellaa frequency irratti moderate, qorannoo hin measure. Nageenya qabu dubbisuu: "yaalaa AI daataa kee leak godha" miti, "iccitii dhuunfaa solved" miti; "istaandaardii iccitii dhuunfaa sakatta'a isa hamaa haalawwan ofii hin argu, vulnerable dhukkubsattoota irratti — sakatta'a jijjiiruu qaba."

# Maddoota

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