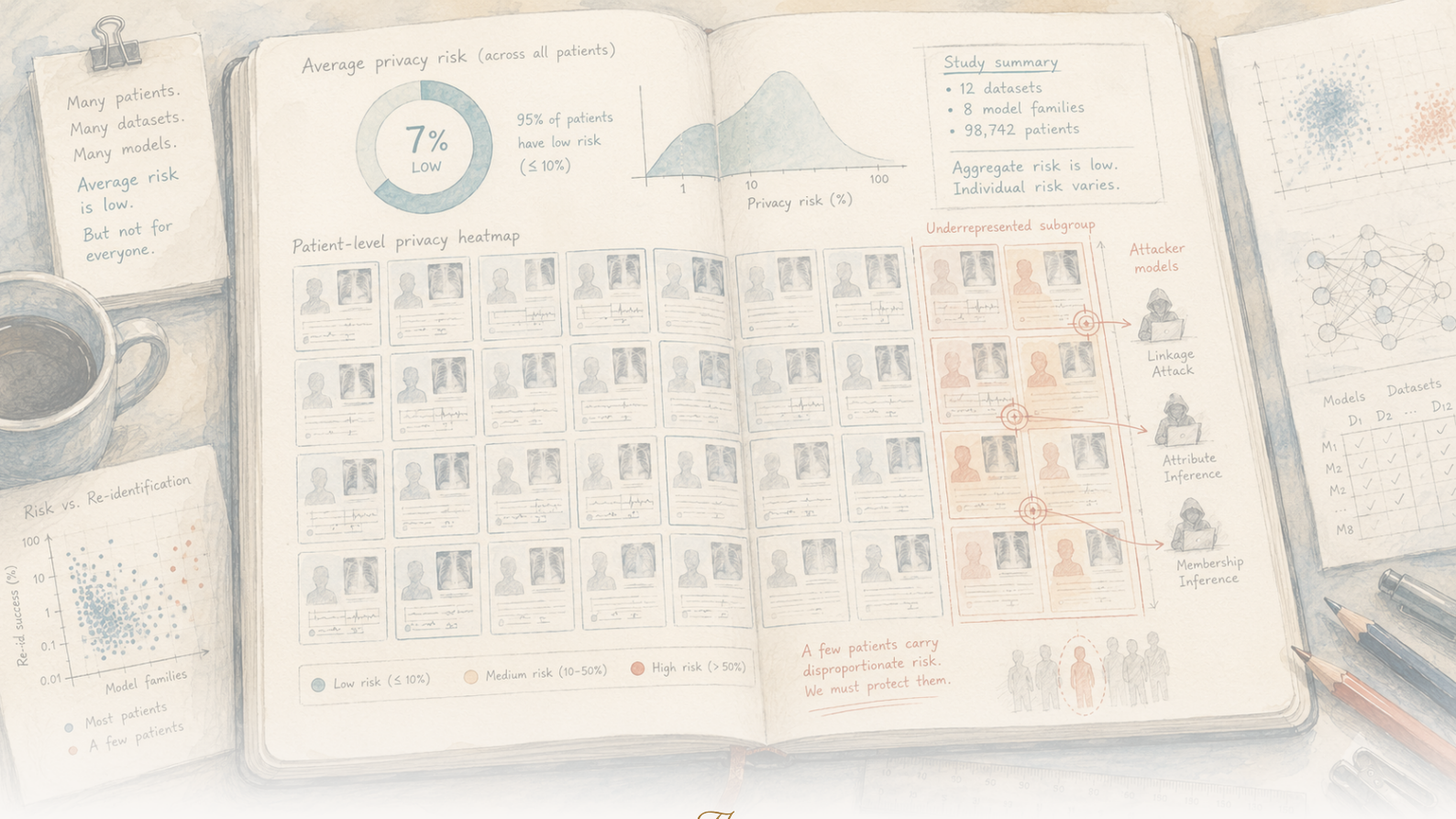


The CLEAN PAPER

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

Medical AI በአማካይ private ሊሆን እና አሁንም የተወሰኑ patientsን ሊያጋልጥ ይችላል — በተለይ underrepresented groupsን

"Privacy-preserving" ተብሎ የሚጠራ medical-AI model ብዙውን ጊዜ በአንድ average number ላይ ይመሰረታል። ይህ ጥናት ያ ቁጥር የተሳሳተው ፈተና ነው ይላል። Membership inference attacks በመመርመር — የአንድ ሰው record model training data ውስጥ ነበር ወይስ አልነበረም የሚያሳዩ እና ስለ disease መረጃ ሊጠቁሙ የሚችሉ — ደራሲዎቹ riskን በaggregate ሳይሆን per patient በሰባት medical datasets እና ብዙ models ላይ ለኩ። Average ላይ safe የሚመስሉ models አንዳንድ individualsን almost perfectly ሊለዩ ይችላሉ (AUC ≥ 0.95)፣ በተለይ underrepresented groups፣ ችግኙም model size ሲጨምር ይባባሳል።

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

የግላዊነት guarantee ለአማካይ ሳይሆን ለእያንዳንዱ ሰው የሚሰጥ ተስፋ ነው

የሕክምና-AI ሞዴል "ግላዊነት-preserving" ተብሎ ሲጠራ ይህ አባባል ብዙውን ጊዜ በአንድ ቁጥር ላይ ይመሰረታል፡- ሞዴሉን ለማሰልጠን መረጃቸው የተጠቀመባቸውን ሁሉንም ታካሚዎች ሲያማክል፣ የአንድ ሰው አባልነት ሊገመት የሚችልበት ዕድል ዝቅ ነው። ይህ አረጋጋጭ ይመስላል። የዚህ ጥናት ጸጥ ያለ ግን አስቸጋሪ ነጥብ፡- የምንመለከተው የተሳሳተውን ቁጥር ነው።

ግላዊነት አማካይ አይደለም። ለእያንዳንዱ ሰው የሚሰጥ ተስፋ ነው — ስልጠና ስብስብ ውስጥ መኖሩ ወደ ኋላ ተመልሶ እንዳያጋልጠው። አማካይ ቁጥር ለ999 ሰዎች ተስፋውን ሊጠብቅ እና ለአንድ ሰው ሙሉ በሙሉ ሊያፈርሰው ይችላል። ተመራማሪዎቹ ግላዊነት አደጋውን ታካሚ በታካሚ ከዚህ በፊት ባልተጠቀመ ዝርዝር ደረጃ ለኩ፣ እና አገኙት ይህን ነው፡- aggregate ላይ safe የሚመስሉ ሞዴሎች የተወሰኑ ሰዎች ስልጠና አባልነትን በአስደናቂ ትክክለኛነት ሊያጋልጡ ይችላሉ — እና የሚጋለጡት በተደጋጋሚ ቀድሞውኑ ያነሰ ጥበቃ ያላቸው ሰዎች ናቸው።

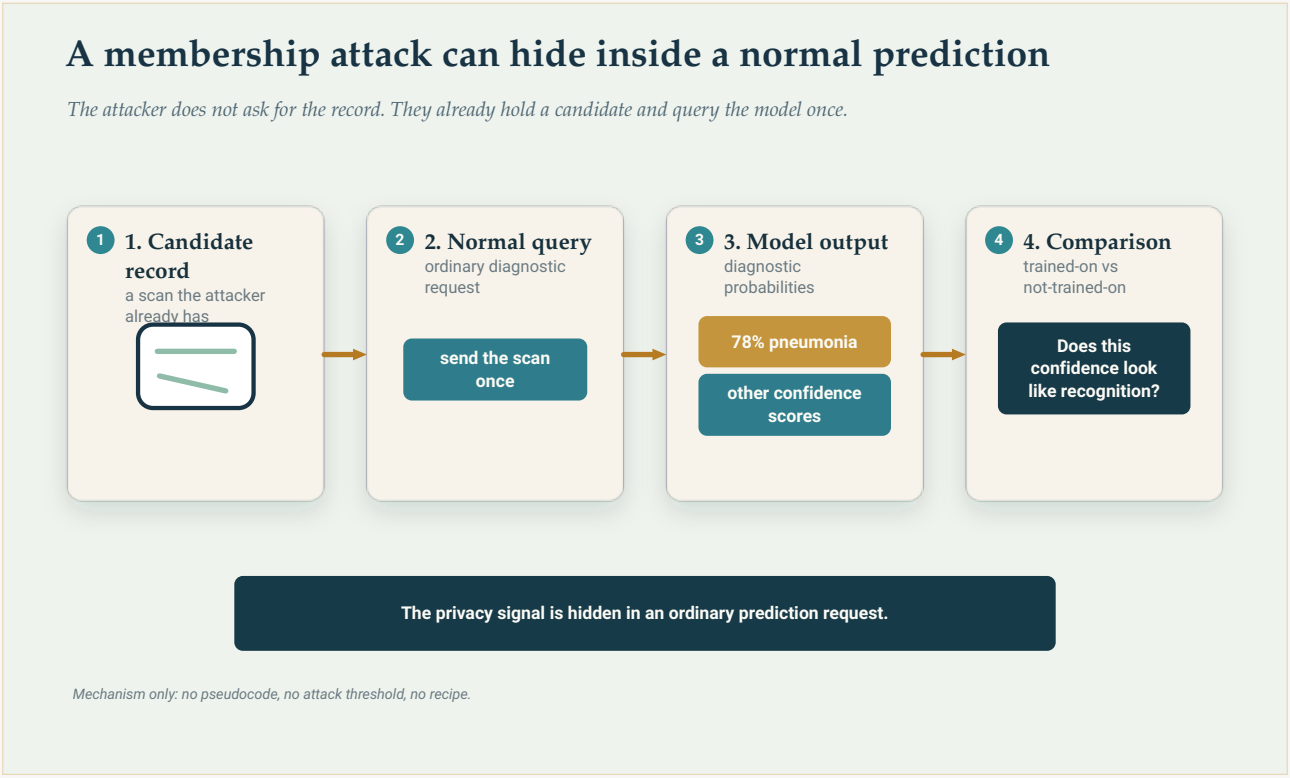
ደራሲዎቹ ምን አደረጉ?

Moritz Knolle **የመራው፣** Daniel Rückert፣ Georg Kaissis **እና ከ**Technical University of Munich፣ Hasso Plattner Institute **እና** Imperial College London **ያሉ ባልደረቦች የታወቀ ግላዊነት** threat **የሆነውን አባልነት መደምደሚያ ጥቃት ወስደው ጥያቄውን ቀየሩ። "ጥቃቱ በየመረጃ ስብስብ አማካይ ምን ያህል ጊዜ ይሳካል?" ከማለት ይልቅ "ይህ የተለየ ታካሚ ምን ያህል የተጋለጠ ነው?" ብለው ጠየቁ።**

Per-ታካሚ ትንተናውን **በሰባት የተመሰረቱ የሕክምና የመረጃ ስብስቦች** ላይ አካሄዱ፣ የመረጃ ዓይነቶቹም የሕክምና ምስል ማንሳት፣ electrocardiograms እና ኤሌክትሮኒክ ጤና recordsን ያካትታሉ። በእያንዳንዱ የመረጃ ስብስብ ላይ 200 ሞዴሎች መርምረዋል። ለእያንዳንዱ ታካሚ በእያንዳንዱ ሞዴል ውስጥ attacker ያ ሰው ስልጠና መረጃ ክፍል ነበር ወይስ አልነበረም በምን ያህል እርግጠኝነት ሊለይ እንደሚችል ገመቱ። ከዚያ ውጤቶችን በበሽታ status፣ self-reported race፣ insurance፣ sex እና ምስል ማንሳት ፕሮቶኮል ከፋፈሉ።

> **አባልነት መደምደሚያ ጥቃት በትክክል ምንድን ነው?**
>
> Leakው "ሞዴሉ የእርስዎን መዝገብ ሙሉ በሙሉ ያወጣል" የሚለው አይደለም። ስለ አባልነት — መረጃዎ ስልጠና ስብስብ ውስጥ ነበር ወይስ አልነበረም — ነው።
> ሞዴል በመደበኛነት ምን ያደርጋል? Scan ትሰጠዋለህ፣ ለምሳሌ chest X-ray፣ እና probabilities ይመልሳል፡- 78% pneumonia፣ እና ሌሎች findings። ጥቃቱ የሚጠቀመው ይህን መደበኛ የምርመራ output ብቻ ነው።
> ድክመቱ ይህ ነው፡- ሞዴል ብዙ ጊዜ በስልጠና ውስጥ በተመለከታቸው ትክክለኛ examples ላይ ከአዲስ examples ይልቅ ትንሽ **ይበልጥ confident** ይሆናል። በፈተናው ከዚህ በፊት ጥያቄዎቹን ያየ ተማሪ እንደሚመስል፣ በእነዚያ ጥያቄዎች ላይ ትንሽ ከመጠን በላይ እርግጠኛ ይሆናል። ያንን ምልክት በመጠቀም አንድ መዝገብ ሞዴሉ ያየው ነበር ወይስ አልነበረም መገመት አባልነት መደምደሚያ ነው።
> ጥቃቱ በአጭሩ፡- attacker የአንድ ሰው scan አለው። ወደ ሞዴል እንደ መደበኛ የምርመራ request አንድ ጊዜ ይልከዋል እና confidence ይመለከታል። ከዚያ ያ confidence በscanው ላይ የተማረ ሞዴሉን ይመስላል ወይስ ያልተማረ ሞዴሉን ብሎ ያነፃፅራል። ይህን reference ሞዴል በራሱ በordinary ኮምፒውተር ማሰልጠን ይችላል። እውነተኛ ሞዴሉ በዚህ scan ላይ ከመጠን በላይ እርግጠኛ ከሆነ፣ "ያውቀዋል" የሚል ምልክት ሊሆን ይችላል።

> የሚያስጨንቀው ነገር ጥያቄው እንደ ጥቃት አይመስልም፤ clinician ሊያደርገው የሚችለው ያው መደበኛ የምርመራ query ነው። ግላዊነት ምልክትው በordinary ትንበያ ውስጥ ተደብቋል። እስካሁን ይህ በተገለጹ ላቦራቶሪ ግምቶች ላይ የታየ እንጂ በእውነተኛ ዓለም ጥቃት ተይዞ የታየ አይደለም።
> መዝገቡ ራሱ ካልወጣ አባልነት ለምን ያስጨንቃል? ምክንያቱም አባልነት ራሱ እውነታ ነው። ሞዴሉ በተወሰነ cancer immunotherapy በተቀበሉ ታካሚዎች ላይ ከተሰለጠነ፣ የእርስዎ መዝገብ በስልጠና ስብስብ ውስጥ ነበር ብሎ ማረጋገጥ ያን cancer እንደነበረዎት ሊጠቁም ይችላል። መዝገብ content ተዘግቶ ይቀራል፤ የ‘‘አባልነት’’ እውነታ ነው የሚወጣው።



ጥቃቱ ልዩ ግላዊነት API አይፈልግም። ሞዴል ቀድሞ ያየውን መዝገብ ላይ ከመጠን በላይ confident ከሆነ፣ መደበኛ የምርመራ query ራሱ አባልነት ምልክት ሊያፈስ ይችላል።

ምን አገኙ?

ሦስት ዋና findings አሉ።

Averages የተጋለጡትን ይደብቃሉ። Aggregate ላይ በመለካት — የተለመደው ዘዴ — ብዙ ሞዴሎች ግላዊነት ጥሩ ይመስላሉ፦ ጥቃቱ ለአብዛኞቹ ታካሚዎች ከchance አይሻልም። Per-ታካሚ እይታ ግን ሌላ ምስል አሳየ። ይህ ልዩነት በጥናት announcement ላይም በግልጽ ተጠቅሷል። አንዳንድ individuals ላይ ጥቃቱ **almost perfectly** ይሳካል፣ ጥቃት AUC **0.95 ወይም ከዚያ በላይ** (0.5 coin toss፣ 1.0 flawless)፣ በየመረጃ ስብስብ-wide average ግን ከchance አይሻልም ብሎ ሊታይ ይችላል።

Average ጥቃት success ወደ chance ቅርብ ሊሆን ይችላል፣ ግን ትንሽ የrecords tail በጣም የተጋለጠ ሊቀር ይችላል። የጥናቱ ነጥብ ግላዊነት ታካሚ-level ላይ መፈተሽ እንዳለበት ነው።

Exposureው እኩል አይደለም — underrepresented groupsን ይመታል። Near-perfect ጥቃት ላይ የበለጠ የተጋለጡ ታካሚዎች በsystematic ሁኔታ መረጃ ውስጥ **underrepresented** የሆኑ groups ነበሩ፦ minority ethnicities፣ rare-በሽታ phenotypes፣ unusual ምስል ማንሳት characteristics። ኤሌክትሮኒክ-መዝገብ የመረጃ ስብስብ ውስጥ ጥቁር ታካሚዎች በmost-vulnerable records መካከል ከተጠበቀው **31% በላይ** ተወክለው ነበር፤ mammography ስብስብ ውስጥ suspicious-for-malignancy scans በዚያ extreme-አደጋ tail **+1,179%** ተጨማሪ ውክልና ነበራቸው። እነዚህ ሁለት ቁጥሮች examples ናቸው፣ ሙሉ ምስሉ አይደሉም። እንዲሁም relative over-representation ናቸው፦ የጥቁር ታካሚዎች 31% ወይም suspicious mammograms 1,179% ተጋልጠዋል ማለት አይደለም። በsmall extreme-አደጋ tail ውስጥ ከተጠበቀው ምን ያህል በላይ እንደሚታዩ ነው። አንድ ሞዴል አንድ rare ታካሚን ከሚመስሉ ሌሎች examples ጋር ለማደብዘዝ ያነሰ መረጃ ስላለው፣ መዝገቡ ይበልጥ ይታያል — ይህም አባልነት ጥቃት በትክክል የሚያገኘው ነው።

ትልቅ ሞዴሎች ችግሩን ያባብሳሉ። Near-perfect ጥቃቶች የተጋለጡ ታካሚዎች ቁጥር ከሞዴል capacity ጋር በፍጥነት ጨመረ። በአንድ dermatology የመረጃ ስብስብ ውስጥ shareው በትንሹ ሞዴል ወደ zero ቅርብ ነበር፣ በትልቁ ሞዴል ግን ወደ **አንድ በአሥር** ደረሰ። የሕክምና AI ሞዴሎች እየበዙ እና እየተጠናከሩ ሲሄዱ ይህ specific አደጋ ይበልጥ ከባድ ሊሆን ይችላል።

“Safe on average” ለምን የተሳሳተ ፈተና ነው?

Low average አደጋውን እንደ clean bill of ጤና ማንበብ ቀላል ነው። የጥናቱ ዋና ትምህርት ግን እዚህ averaging ትንሽ ያልትክክል ብቻ ሳይሆን የተሳሳተውን ነገር እየለካ ነው።

ግላዊነት guarantee በአማካይ ሰው ላይ ሳይሆን በጣም የተጋለጠው ሰው ላይ መጠበቅ አለበት። 1,000 ታካሚዎች ውስጥ 999 መለየት ካልተቻለ እና አንዱ ግን almost perfectly ከተለየ፣ ለዚያ አንድ ሰው ሞዴሉ “99.9% private” ነው ማለት ትርጉም የለውም። እና ያ አንድ ሰው በተገመተ ሁኔታ rare-በሽታ ታካሚ ወይም ethnic-minority ታካሚ ከሆነ፣ ሜትሪክው incomplete ብቻ ሳይሆን discriminatory የሆነ blind spot አለው።

ጥናቱ የሚገድደን ለውጥ ይህ ነው፡- “ሞዴሉ በአማካይ ምን ያህል private ነው?” ከማለት ወደ “በጣም የተጋለጠው ታካሚ ማነው፣ እና ይህ ሰው ማንነቱ ምንድን ነው?” መሄድ።

ይህ ምን አያረጋግጥም — ወይም ምን አይናገርም?

- የሕክምና AI መተው አለብን ብሎ አይልም። ደራሲዎቹ mitigationን ይጠይቃሉ፣ retreatን አይደለም።
- እያንዳንዱ የሕክምና ሞዴል leak ያደርጋል ወይም የተዘረጋ ሞዴል ተጠቅቶበታል ብሎ አያሳይም። አደጋ እንዳለ እና እኩል እንዳልተከፋፈለ ያሳያል።
- Recordsዎ አሁን ተጋልጠዋል ብሎ አያሳይም። ጥቃቱ access to ሞዴል፣ እጩ መዝገብ እና attacker reference infrastructure ይፈልጋል።
- ታካሚ *content* እንደሚወጣ አያሳይም። የሚወጣው አባልነት እውነታ ነው።
- Disparityን ወደ አንድ headline ቁጥር አያሳንስም። Strong reproducible ውጤት patternው ነው፡- aggregate metrics individual አደጋውን ያነሱ ያሳያሉ፣ underrepresented ታካሚዎችም ዋና ጫናውን ይሸከማሉ።

ማስረጃው ምን ያህል ጠንካራ ነው?

ይህ empirical methodological ውጤት ነው፣ እና ጠንካራ ነው። Aggregate ግላዊነት metrics per-ታካሚ አደጋውን እንደሚያነሱ የሚያሳዩ፣ residual አደጋ underrepresented groups ላይ እንደሚሰበሰብ፣ ሞዴል capacity ሲጨምር እንደሚባባስ — ይህ pattern በሰባት የተለያዩ መረጃ-type የመረጃ ስብስቦች እና ብዙ ሞዴሎች ላይ ተደግሟል። ይህ breadth ውጤቱን ከsingle-የመረጃ ስብስብ artefact ይለየዋል።

ሁለት honest caveats አሉ። መጀመሪያ፣ ይህ አደጋ *demonstration* ነው፣ ደራሲዎቹ በገነቡት ጥቃቶች መሠረት capable attacker በተገለጹ ግምቶች ምን ያህል ሊሳካ እንደሚችል ይለካል። በእውነተኛ world ምን ያህል ጊዜ ጥቃቶች እንደሚከሰቱ አይለካም። ሁለተኛ፣ near-perfect success የሚሉ አስደናቂ ቁጥሮች በትርጉም በጣም የተጋለጡ individualsን ይመለከታሉ። የጥናቱ ነጥብ በትክክል ይህ ነው፡- tailው average ከሚያሳየው ብዙ ከባድ ነው።

ተገቢው stance alarm ወይም dismissal አይደለም። መደበኛ የሕክምና-AI ግላዊነት certification የworst ጉዳዮችን እንደማያይ እና እነዚያ worst ጉዳዮች በትንሹ ማርጂን ባላቸው ታካሚዎች ላይ እንደሚወድቁ የሚያሳይ ጥንቃቄ ያለው multi-የመረጃ ስብስብ ጥናት ነው።

ለምን ይጠቅማል?

ሁለት ነገሮች ይህን ከtechnical footnote ያሳድጉታል።

መጀመሪያ፣ “ግላዊነት-preserving” ቃል ምን ማለት እንዲችል ይቀይራል። ሞዴል በaverage ግላዊነት ነጥብ ብቻ ከተለቀቀ፣ ነጥብው በእውነት ዝቅ ሊሆን ይችላል እና አሁንም near-perfect አባልነት ጥቃት የሚቻልባቸው ታካሚዎችን ሊደብቅ ይችላል። የጥናቱ ተግባራዊ demand ግልጽ ነው፡- release ከመደረጉ በፊት ግላዊነት አደጋውን per ታካሚ መገምገም፣ deployed ሞዴሎች accessን መቆጣጠር፣ እና differential ግላዊነት ያሉ techniquesን መጠቀም። Differential ግላዊነት በስልጠና ውስጥ በጥንቃቄ የተመጠነ noise በመጨመር አባልነት ጥቃቶችን ያዳክማል፣ ግን ሞዴል utility ላይ በእውነት cost አለው፤ ግላዊነት–utility trade-off ነፃ አይደለም። “Averageን ፈትነናል” ብቻ እንደ full ግላዊነት check መቆጠር አይገባም።

ሁለተኛ፣ ግላዊነትን ከfairness ጋር ያገናኛል። የሕክምና መረጃ ውስጥ underrepresented የሆኑ groups — እና ስለዚህ የሕክምና AI ቀድሞውኑ ያነሰ ጥሩ ሊያገለግላቸው የሚችሉ — እነዚያው ሞዴል ግላዊነት ያነሰ የሚጠብቃቸው ሰዎች መሆናቸው ታይቷል። ሁለቱ ችግሮች በተለያዩ departments የሚያዙ አይደሉም፤ ተመሳሳይ ሰዎችን ይመለከታሉ።

የ“average አደጋ ዝቅ ነው፣ ስለዚህ ግላዊነት ጥሩ ነው” ታሪክን ጥናቱ ይፈትነዋል። ቴክኖሎጂውን ለማስፈራራት አይደለም፤ መደበኛን ወደሚገባው ቦታ ለመውሰድ ነው፡- በጣም ሊጎዳ የሚችለውን ሰው ካልፈተሸክ ግላዊነት promise ጠብቄዋለሁ ማለት አትችልም።

ንጹህ ማጠቃለያ

አባልነት መደምደሚያ ጥቃቶች የአንድ ሰው መዝገብ AI ሞዴል ስልጠና መረጃ ውስጥ ነበር ወይስ አልነበረም ለመወሰን ይሞክራሉ። አባልነት ራሱ ስለተወሰነ በሽታ ወይም ሕክምና መረጃ ሊጠቁም ስለሚችል፣ መዝገቡ ሳይወጣም ይህ genuine ግላዊነት leak ነው። ቀደም ያሉ ሥራዎች ጥቃት successን የመረጃ ስብስብ-wide *average* ላይ ይለኩ ነበር። ይህ ጥናት ግን አደጋውን *per* ታካሚ ለካ፣ በሰባት የሕክምና የመረጃ ስብስቦች — ምስል ማንሳት፣ ECG፣ ኤሌክትሮኒክ records — እና ብዙ ሞዴሎች ላይ። ውጤቱ፡- aggregate metrics አደጋውን በጣም ያነሱ ያሳያሉ፤ አንዳንድ individuals almost perfectly ሊለዩ ይችላሉ (ጥቃት AUC $\geq$ 0.95) እንኳን የመረጃ ስብስብ-wide average safe ቢመስልም። Most vulnerable ታካሚዎች በsystematic ሁኔታ underrepresented groups — minority ethnicity፣ rare በሽታ፣ unusual ምስል ማንሳት — ናቸው፣ ችግኙም ሞዴል size ሲጨምር ይባባሳል። ደራሲዎቹ የሕክምና AIን እንተው አይሉም፤ ግላዊነትን individual level ላይ እንለካ፣ ሞዴል accessን እንቆጣጠር እና differential ግላዊነት እንጠቀም ይላሉ።

ያለ ማጋነን ማረጋገጫ

ጥናቱ የሚያሳየው: በሰባት የሕክምና የመረጃ ስብስቦች እና ብዙ ሞዴሎች ላይ per-ታካሚ አባልነት-መደምደሚያ አደጋ ለአንዳንድ individuals aggregate metrics ከሚያሳዩት በጣም ከፍ ነው — እስከ near-perfect ጥቃት success (AUC $\geq$ 0.95) — residual አደጋ underrepresented groups ላይ በብዛት ይወድቃል፣ ሞዴል capacity ሲጨምርም ይጨምራል።

ሊሆን የሚችል ግን ያልተረጋገጠ: እውነተኛ-world attackers አሁን deployed ክሊኒካዊ ሞዴሎች ላይ ይህን እየጠቀሙ እንደሆነ። ጥናቱ capability እና ስርጭትን በተገለጹ ግምቶች ያሳያል፣ frequency in the wildን አይለካም።

የማያሳየው: የሕክምና AI መተው እንዳለብን፤ እያንዳንዱ ሞዴል leaks እንዳለው፤ specific deployed ሞዴል ተጥሷል እንደሆነ፤ ታካሚ መዝገብ contents እንደሚወጡ፤ ወይም disparity በአንድ single ቁጥር እንደሚጠቃለል አያሳይም።

ዋና ገደቦች: Worst-ጉዳይ አደጋ በደራሲዎቹ ጥቃቶች ይለካል፣ የታየ incidents አይደሉም፤ ከፍተኛ ቁጥሮቹ most-

exposed individualsን በትርጉም ይገልጻሉ፤ exact per-group disparities ከአንድ ቁጥር ይልቅ በየመረጃ ስብስቦች ላይ consistent pattern ናቸው።

አጠቃላይ አንባቢ ምን ያህል ሊተማመን ይገባል? Aggregate ግላዊነት metrics individual አደጋውን እንደሚያነሱ ያሳዩ፣ residual አደጋ underrepresented ታካሚዎች ላይ እንደሚሰበሰብ፣ እና ሞዴል size ሲጨምር ችግኙ እንደሚባባስ ላይ ከፍተኛ እምነት። እውነተኛ-world ጥቃት frequency ላይ መካከለኛ እምነት ብቻ፣ ምክንያቱም ጥናቱ አይለካውም። ትክክለኛው ንባብ "የሕክምና AI መረጃዎን ያፈስሳል" አይደለም፣ "ግላዊነት ተፈቷል"ም አይደለም። የተለመደው ግላዊነት check የworst ጉዳዮችን አያይም፣ እና እነዚያ ጉዳዮች በተጋለጡ ታካሚዎች ላይ ይወድቃሉ — ስለዚህ checkው መቀየር አለበት።

ምንጮች

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