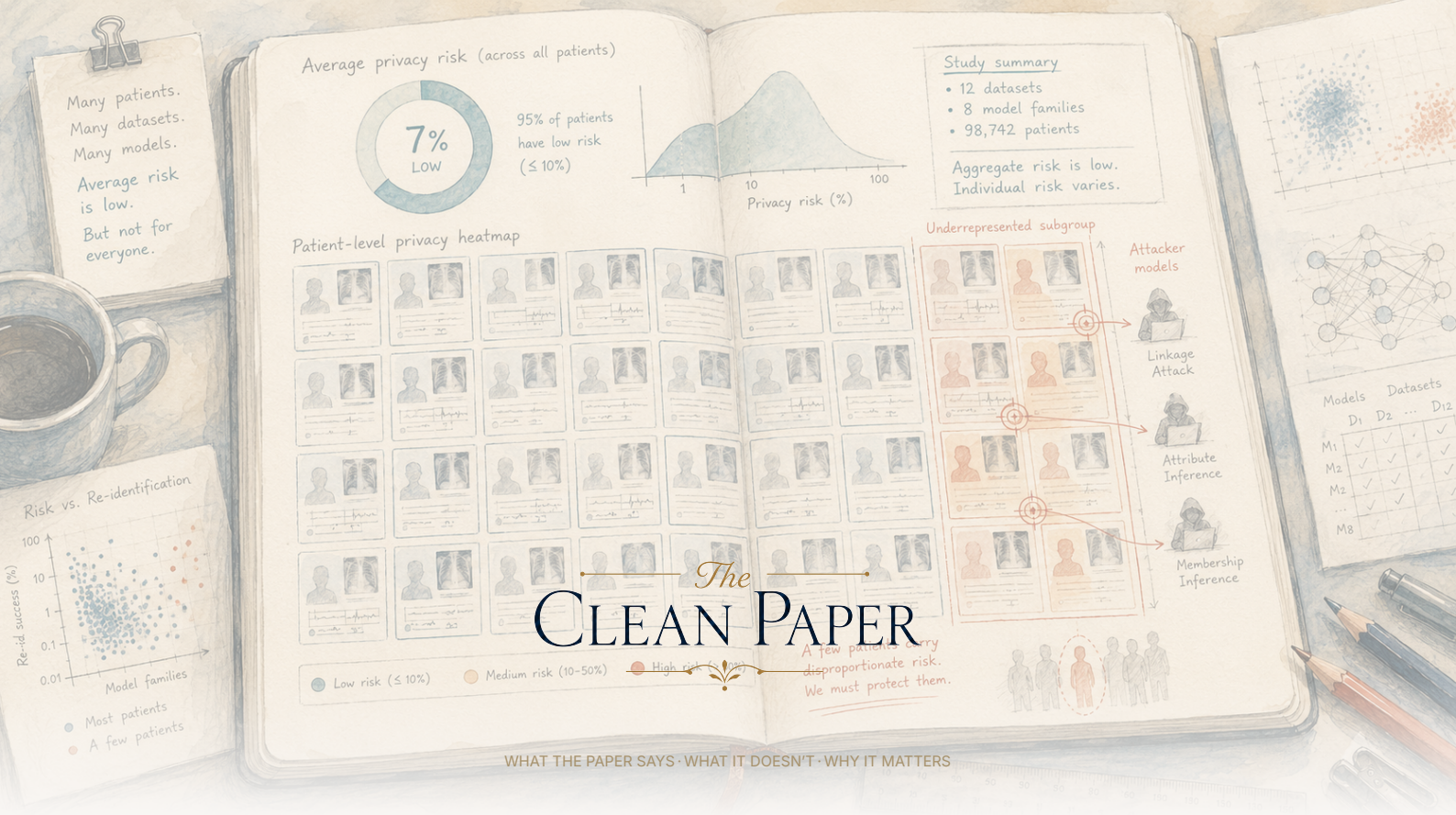


Медицинский ИИ может быть приватным в среднем и одновременно раскрывать отдельных пациентов — особенно недопредставленных

Когда медицинскую модель ИИ называют «сохраняющей приватность», это обычно основано на одном среднем показателе. Это исследование показывает, что такая проверка неверна. Изучая атаки на определение принадлежности к обучающей выборке — они позволяют выяснить, была ли запись конкретного человека в обучающих данных модели и тем самым могут раскрыть, что у человека было определённое заболевание, — авторы измеряли риск не в совокупности, а для каждого пациента отдельно, на семи медицинских датасетах (изображения, ЭКГ, электронные записи) и множестве моделей. Картина такова: модели, безопасные по среднему показателю, иногда позволяют почти безошибочно идентифицировать конкретных людей (AUC атаки $\geq 0{,}95$); среди раскрываемых систематически преобладают недопредставленные группы — этнические меньшинства, редкие болезни, необычные изображения; а с увеличением моделей проблема усиливается. Авторы не призывают отказаться от медицинского ИИ — они предлагают измерять приватность для каждого пациента, контролировать доступ к моделям и применять дифференциальную приватность. Неудобная суть: «приватна в среднем» — не гарантия приватности, и чаще всего она подводит тех пациентов, которые и без того защищены хуже всего.

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

Гарантия приватности — это обещание человеку, а не среднему значению

Когда о медицинской модели ИИ говорят, что она «сохраняет приватность», обычно это утверждение опирается на одно число: среди всех пациентов, на данных которых обучали модель, средняя вероятность определить, входили ли данные конкретного человека в обучающий набор, невелика. Звучит успокаивающе. Тихий и неудобный вывод этой работы состоит в том, что это просто не то число, которое нужно смотреть.

Приватность — не среднее значение. Это обещание каждому отдельному человеку: участие его данных в обучающем наборе не должно впоследствии его раскрыть. А средний показатель может выполнять это обещание почти для всех и одновременно полностью нарушать его для нескольких людей. Исследователи измерили риск приватности для каждого пациента отдельно, с разрешением, которого раньше не применяли, и увидели именно это: модели, которые в совокупности выглядят безопасными, могут почти безошибочно выдавать сам факт участия конкретных людей в обучающих данных — и непропорционально часто это именно те люди, которые и без того защищены хуже всего.

Что сделали авторы

Команда под руководством Морица Кнолле вместе с Даниэлем Рюккертом (оба — Technical University of Munich) и Георгом Кайссисом (Hasso Plattner Institute), а также коллегами из Imperial College London взяла хорошо известную угрозу приватности — **атаку на определение принадлежности к обучающей выборке (membership inference attack)** — и изменила вопрос. Вместо «как часто эта атака в среднем успешна по всему набору данных?» они спросили: «насколько уязвим *этот конкретный пациент*?»

Такой анализ на уровне отдельного пациента они провели на **семи устоявшихся медицинских наборах данных**, охватывающих очень разные типы информации — медицинские изображения, электрокардиограммы и электронные медицинские записи, — и для каждого исследовали по 200 моделей. Для каждого пациента в каждой модели они оценивали, насколько уверенно злоумышленник мог бы определить, была ли карта этого человека частью обучающих данных, а затем разбивали результаты по группам: статусу заболевания, самоидентифицированной расовой принадлежности, страховке, полу и протоколу визуализации.

> **Что на самом деле представляет собой атака на определение принадлежности к обучающей выборке**
>
> Утечка здесь тоньше, чем «модель выдаёт вашу медицинскую карту». Речь

идёт о *принадлежности* — самом факте того, что ваши данные находились в обучающем наборе.

Начнём с того, что такая модель делает в обычном режиме. Вы подаёте ей изображение — скажем, рентген грудной клетки — и получаете вероятности: 78% вероятности пневмонии, а также оценки для кардиомегалии, отёка, консолидации. Именно этот повседневный диагностический ответ и использует атака. Ничего экзотического.

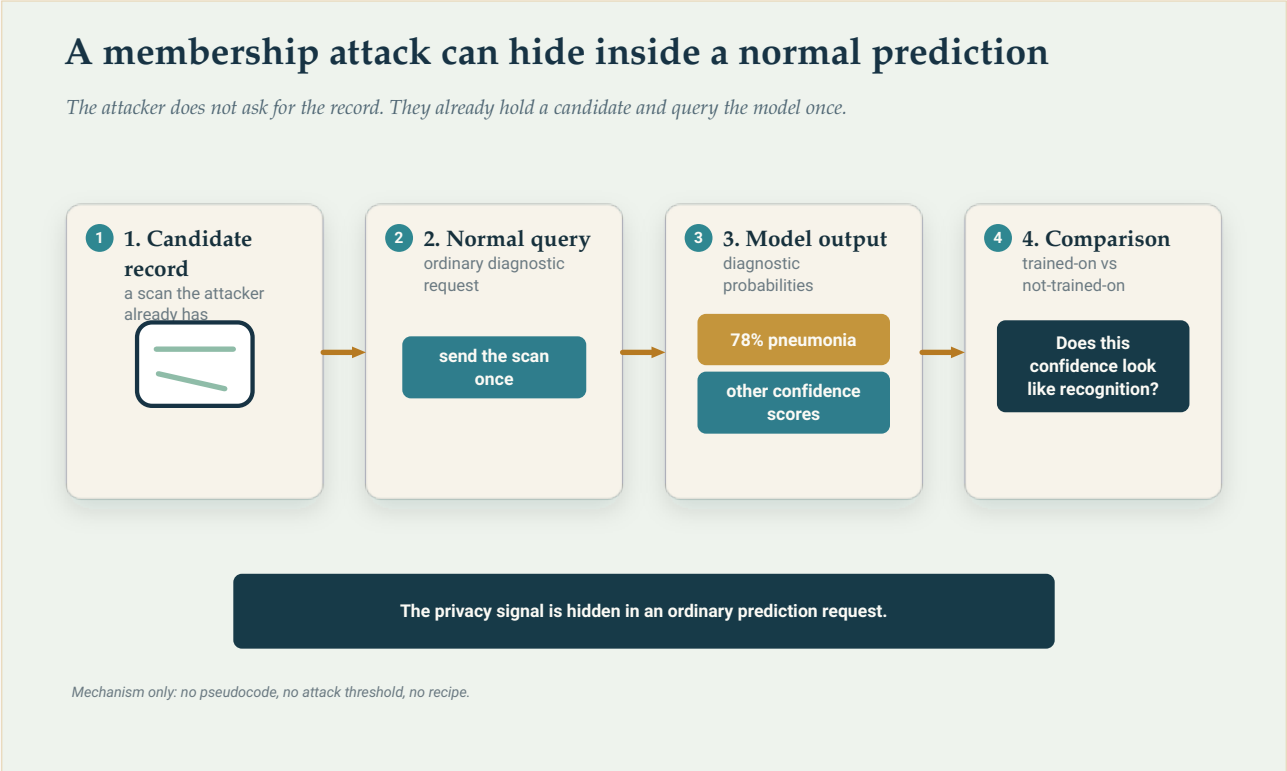
Слабое место заключается в следующем: модель обычно немного **увереннее** именно на тех примерах, на которых её обучали, чем на тех, которых она никогда не видела. Представьте студента, который тайком заранее видел экзаменационные вопросы: на уже отработанные вопросы ответы приходят чуть быстрее и чуть увереннее. Определить по этому признаку, была ли конкретная запись среди примеров, которые модель «изучала», — это и есть атака определения принадлежности к обучающей выборке.

Итак, атака выглядит так. Кто-то хочет узнать, использовался ли скан конкретного человека для обучения модели. Он **не** просит модель отдать запись — у него уже есть скан-кандидат, собственный скан этого человека или очень близкая копия. Он один раз подаёт его как обычный диагностический запрос и смотрит, насколько уверенно отвечает модель. Затем сравнивает эту уверенность с поведением модели, которая *обучалась* на таком скане, и модели, которая *не обучалась* на нём. Такое сравнение можно недорого построить, обучив собственную вспомогательную «эталонную модель», чтобы она показала характерную избыточную уверенность; для этого достаточно обычного компьютера без специального оборудования. Если настоящая модель необычно уверена именно на этом скане — словно узнаёт его, — это сильный сигнал, что скан входил в обучающие данные.

Самое неприятное то, что сам запрос совсем не выглядит как атака: это тот же диагностический запрос, который мог бы сделать клиницист, а сигнал о приватности спрятан внутри обычного прогноза. Поэтому риск вполне конкретен — хотя пока его демонстрировали при чётко определённых лабораторных предположениях, а не наблюдали в реальных атаках.

Зачем вообще важна принадлежность, если сама запись никогда не утекает? Потому что *принадлежность тоже является фактом*. Если модель обучали на пациентах, получавших определённую иммунотерапию рака, то подтверждение того, что ваша запись входила в этот набор, может раскрыть, что у вас, вероятно, был этот рак — именно тот тип информации, который страховщик или работодатель не должен иметь возможности вывести. Содержимое записи остаётся закрытым; наружу выходит факт принадлежности к группе.

Авторы прямо описывают предпосылки атаки — доступ к обычным прогнозам модели, наличие записи-кандидата и собственной эталонной модели злоумышленника — не как инструкцию, а для того, чтобы угрозу можно было анализировать конкретно, а не отмахиваться от неё.



Для атаки не нужен специальный API приватности. Обычный диагностический запрос может нести сигнал о принадлежности к обучающей выборке, если модель необычно уверена на записи, которую уже видела.

Что они обнаружили

Три результата, каждый острее предыдущего.

Средние значения скрывают самых уязвимых. Если измерять в совокупности — как обычно и делают, — многие модели выглядят достаточно приватными: для большинства пациентов атака работает не лучше случайного угадывания. Но на уровне отдельного человека картина иная. Как Кнолле сказал в сообщении об исследовании, предыдущие оценки «всегда измеряли лишь средний риск для всех пациентов. Мы впервые исследовали риск на уровне отдельных пациентов — и картина оказалась совершенно иной». Для некоторых людей атака работает **почти идеально** — авторы измеряют это как AUC атаки **0,95 или выше** (0,5 — случайное угадывание, 1,0 — безошибочный результат), даже когда средний показатель по всему набору данных не лучше случайности.

Средний риск может оставаться близким к случайному уровню, тогда как небольшой «хвост» записей раскрывается намного сильнее. Главный тезис статьи: приватность нужно проверять на уровне отдельного пациента, а не только в совокупности.

Риск распределён неравномерно — и приходится на недопредставленных. Пациенты, наиболее уязвимые для почти безошибочной атаки, систематически принадлежали к группам, **недопредставленным в данных**: этническим меньшинствам, редким фенотипам заболеваний, необычным характеристикам изображений. В наборе электронных медицинских записей темнокожие пациенты встречались среди наиболее уязвимых записей **на 31% чаще**, чем ожидалось; в наборе маммографии сканы, помеченные как подозрительные на злокачественность, были сверхпредставлены в этой зоне риска на поразительные **+1 179%**. (Эти два числа — примеры, а не вся картина: статья описывает тот же перекос для нескольких групп. И это *относительное* сверхпредставление внутри малого «хвоста» крайнего риска — то есть насколько чаще такие пациенты встречаются среди наиболее раскрываемых записей, а не доля всех темнокожих пациентов или всех пациентов с подозрительными маммограммами, которые раскрываются.) У модели меньше похожих примеров, среди которых такие пациенты могли бы «затеряться», поэтому их записи сильнее выделяются — а именно это и улавливает такая атака. Провал приватности не случаен; он концентрируется на людях, которые уже находятся на периферии данных.

Более крупные модели ухудшают ситуацию. Число пациентов, уязвимых для почти идеальных атак, резко росло вместе с ёмкостью модели. В одном дерматологическом наборе данных доля поднялась от практически нуля у самой маленькой модели примерно до **одного из десяти** у самой большой. Поскольку медицинские модели ИИ

становятся крупнее и мощнее — а именно в этом направлении движется вся отрасль, — этот конкретный риск, предупреждают авторы, усиливается, а не ослабевает.

Вывод Рюккерта резок: «Это неприемлемый риск. Медицинские данные чрезвычайно чувствительны».

Почему «безопасна в среднем» — неправильная проверка

Есть соблазн прочитать низкий средний риск как справку «всё в порядке». Главный урок работы состоит в том, что усреднение здесь не просто неточно — оно измеряет не то.

Гарантия приватности имеет смысл только тогда, когда действует для человека с наибольшим риском, а не для человека в середине распределения. Модель, в которой 999 из 1 000 пациентов невозможно идентифицировать, а одного можно определить почти идеально, не является «приватной на 99,9%» ни в каком смысле, имеющем значение для этого одного человека. А если этим человеком предсказуемо оказывается пациент с редкой болезнью или представитель этнического меньшинства, метрика не просто неполна — она незаметно дискриминирует. Она измеряет безопасность большинства и называет это безопасностью для всех.

Именно такого сдвига требует статья: от *насколько приватна эта модель в среднем?* к *кто является наиболее раскрываемым пациентом и к какой группе он относится?* Это разные вопросы, и только второй по-настоящему касается гарантии приватности.

Чего эта работа *не* доказывает — и не утверждает

- Она **не** говорит, что от медицинского ИИ следует отказаться. Рамка авторов — снижение риска, а не отступление: измерять и исправлять проблему, а не прекращать разработку.
- Она **не** означает, что каждая медицинская модель что-то утечёт или что какую-либо конкретную развёрнутую модель уже атаковали. Она показывает, что *риск существует и распределён неравномерно*, а стандартные агрегированные метрики его пропускают.
- Она **не** означает, что ваши записи уже раскрыты. Атака требует конкретных условий — доступа к модели, записи-кандидата и собственной инфраструктуры злоумышленника, — а не случайной возможности любого человека.
- Она **не** показывает утечку *содержимого* карты пациента. Утекает принадлежность — сам факт включения в обучающую выборку. Это опасно по другой причине, а не потому, что модель просто выдаёт запись.
- Она **не** сводит неравенство к одному громкому числу. Сильный и воспроизводимый результат — это *паттерн*: агрегированные метрики занижают индивидуальный

риск, а наибольший риск несут недопредставленные пациенты — в разных наборах данных и сотнях моделей.

Насколько убедительны доказательства?

Это эмпирический методологический результат, и он достаточно прочен. Картина — агрегированные показатели приватности систематически занижают риск для отдельного пациента, остаточный риск концентрируется на недопредставленных группах и растёт с ёмкостью модели — повторилась на **семи наборах данных разных типов** и большом количестве моделей для каждого. Именно такая широта делает утверждение о проблеме измерения убедительным, а не артефактом одного набора данных.

Есть две честные оговорки. Во-первых, это демонстрация *риска*, измеренного атаками, которые построили сами авторы. Она характеризует, насколько успешным *мог бы* быть компетентный злоумышленник при заданных предположениях, а не то, как часто такие атаки действительно происходят. Во-вторых, яркие числа — почти идеальный успех атаки для некоторых пациентов — намеренно описывают людей с худшим результатом. В этом и состоит смысл работы, и читать эти числа нужно как «хвост распределения гораздо тяжелее, чем подсказывает среднее», а не как «большинство пациентов раскрыты».

Правильная реакция — ни паника, ни отмахивание: это тщательное исследование на множестве наборов данных, показывающее, что стандартный способ сертифицировать приватность медицинского ИИ слеп к собственным худшим случаям — и что эти случаи приходятся именно на пациентов, у которых и так меньше всего запаса безопасности.

Почему это важно

Есть две причины, почему это больше, чем техническое замечание.

Во-первых, работа меняет то, что вообще имеет право называться «сохранением приватности». Если модель выпускают со средним показателем приватности, этот показатель может быть действительно низким и при этом скрывать подгруппу пациентов, которых такая атака почти безошибочно идентифицирует. Практическое требование статьи конкретно: перед релизом оценивать риск приватности **для каждого пациента**, контролировать, кто имеет доступ к развёрнутым моделям, и использовать методы вроде **дифференциальной приватности** — небольшой, тщательно откалиброванный шум во время обучения, который снижает эффективность таких атак ценой реального и управляемого снижения полезности модели (компромисс между приватностью и полезностью здесь явный, а не бесплатный). «Мы проверили

среднее» больше не должно считаться достаточной проверкой.

Во-вторых, приватность здесь переплетается со справедливостью. Те же группы, которые недопредставлены в медицинских данных — и поэтому медицинский ИИ уже обслуживает их хуже, — оказываются теми, чью приватность он защищает слабее всего. Отрасль, которая всерьёз работает над первым неравенством, не может считать второе чужой проблемой. Речь идёт о тех же людях.

Успокаивающая история о приватности медицинского ИИ — *мы измерили, средний риск низок, значит всё хорошо* — именно та история, которую эта статья разбирает. Не для того, чтобы отпугнуть от технологии, а чтобы переместить стандарт туда, где ему место: это обещание, о выполнении которого можно говорить лишь после проверки для человека, которому с наибольшей вероятностью может быть причинён вред.

Краткий итог

Атаки на определение принадлежности к обучающей выборке пытаются выяснить, входила ли запись конкретного человека в обучающие данные модели ИИ. Поскольку сама принадлежность может раскрыть чувствительную информацию — например, что у человека было определённое заболевание, — это настоящая утечка приватности даже тогда, когда сама медицинская запись никогда не появляется наружу. Раньше такие атаки оценивали по тому, как часто они успешны *в среднем* по набору данных. В этой работе риск измеряли *для каждого пациента* на семи медицинских наборах данных (изображения, ЭКГ, электронные записи) и множестве моделей для каждого. Оказалось, что агрегированные метрики сильно занижают риск: некоторых людей можно идентифицировать почти безошибочно (AUC атаки $\geq 0{,}95$), даже когда средний показатель по набору данных выглядит безопасным; наиболее уязвимы систематически представители недопредставленных групп (этнические меньшинства, редкие болезни, необычные изображения); а проблема усиливается по мере увеличения моделей. Авторы не призывают отказаться от медицинского ИИ — они предлагают измерять приватность на индивидуальном уровне, контролировать доступ к моделям и применять дифференциальную приватность. Вывод: «приватна в среднем» — не гарантия приватности, а провал этой гарантии чаще всего приходится на тех, кто уже защищён хуже всего.

Проверка без прикрас

Что показывает статья: На семи медицинских наборах данных и множестве моделей для каждого риск определения принадлежности к обучающей выборке для отдельных пациентов гораздо выше, чем подсказывают агрегированные метрики, — вплоть до почти идеального успеха атаки (AUC $\geq 0{,}95$). Этот остаточный риск непропорционально

приходится на недопредставленные группы и растёт с ёмкостью модели.

Что правдоподобно, но не доказано: Что реальные злоумышленники уже используют это против развёрнутых клинических моделей. Исследование демонстрирует *возможность и распределение риска* при заданных предположениях, а не частоту атак в реальном мире.

Чего оно не показывает: Что от медицинского ИИ следует отказаться; что каждая модель имеет утечку или какую-либо конкретную развёрнутую модель уже взломали; что утекает *содержимое* карты пациента, а не факт принадлежности; одного универсального числа неравенства — надёжный результат здесь паттерн, а не одна цифра.

Главные ограничения: Работа измеряет риск худшего случая с помощью атак самих авторов, а не наблюдаемые инциденты; драматические числа намеренно описывают наиболее раскрываемых людей; точные различия между группами показаны как последовательная картина в разных наборах данных, а не сведены к одному числу.

Какой уверенности заслуживает вывод для широкого читателя? Высокой в том, что агрегированные метрики приватности занижают индивидуальный риск, что остаточный риск концентрируется на недопредставленных пациентах и усиливается с увеличением моделей — это показано на многих наборах данных и моделях. Умеренной относительно реальной частоты таких атак, которую это исследование не измеряет. Осторожный вывод: не «медицинский ИИ сливает ваши данные» и не «проблема приватности решена», а «стандартная проверка приватности слепа к собственным худшим случаям, и эти случаи приходятся на наиболее уязвимых пациентов — значит, проверку нужно изменить».

Источники

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