nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00858-y

Reconciling privacy and accuracy in AI for medical imaging

Received: 22 November 2023

Accepted: 20 May 2024

Published online: 21 June 2024

Check for updates

Alexander Ziller[1] ✉, Tamara T. Mueller[1], Simon Stieger[1,2], Leonhard F. Feiner[1,3], Johannes Brandt[1], Rickmer Braren[1,3,4], Daniel Rueckert[1,5] & Georgios Kaissis[1,2,3,5]

Artificial intelligence (AI) models are vulnerable to information leakage of their training data, which can be highly sensitive, for example, in medical imaging. Privacy-enhancing technologies, such as differential privacy (DP), aim to circumvent these susceptibilities. DP is the strongest possible protection for training models while bounding the risks of inferring the inclusion of training samples or reconstructing the original data. DP achieves this by setting a quantifiable privacy budget. Although a lower budget decreases the risk of information leakage, it typically also reduces the performance of such models. This imposes a trade-off between robust performance and stringent privacy. Additionally, the interpretation of a privacy budget remains abstract and challenging to contextualize. Here we contrast the performance of artificial intelligence models at various privacy budgets against both theoretical risk bounds and empirical success of reconstruction attacks. We show that using very large privacy budgets can render reconstruction attacks impossible, while drops in performance are negligible. We thus conclude that not using DP at all is negligent when applying artificial intelligence models to sensitive data. We deem our results to lay a foundation for further debates on striking a balance between privacy risks and model performance.

The rapid rise of artificial intelligence (AI) applications in medicine promises to transform healthcare, offering improvements ranging from specific applications, such as more precise pathology detection or outcome prediction, to the promise of general medical AI[1–5]. However, recent results highlight a substantial vulnerability: AI models may disclose details of their training data. This can happen either inadvertently or be forced through attacks by malicious third parties, also called adversaries. Among the most critical attacks are data reconstruction attacks, where the adversary attempts to extract training data from the model or its gradients[6–17]. Such attacks harbour distinct risks. On one hand, a successful data reconstruction attack severely undermines the trust of patients whose data are exposed. This not only jeopardises the relationship between medical practitioners and patients, but probably also diminishes the willingness of patients to make their health data for the training of AI models or for other research purposes available. This is problematic since the success of AI models in medicine is dependent on the availability of large and diverse real-world patient datasets. On the other hand, a successful attack can also constitute a breach of patient data privacy regulations.

While privacy laws vary globally, the protection of health data is generally considered of high importance. For example, the European Union's General Data Protection Regulation declares the protection of personal data as a fundamental right. Notably, some of these laws deem the removal of personal identifiers (for example, name or

[1] Artificial Intelligence in Healthcare and Medicine, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany. [2] Institute of Machine Learning in Biomedical Imaging, Helmholtz Munich, Neuherberg, Germany. [3] Institute for Diagnostic and Interventional Radiology, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany. [4] German Cancer Consortium (DKTK), Munich partner site, Heidelberg, Germany. [5] Department of Computing, Imperial College London, London, UK. ✉e-mail: alex.ziller@tum.de

date of birth)—de-identification—sufficient protection. However, it has been demonstrated on several occasions that commonly used de-identification techniques such as anonymization, pseudonymization or *k*-anonymity are vulnerable to re-identification attacks[18–20]. This also holds true in the case of medical imaging data. For example, the facial contours of a patient can be obtained from a reconstructed magnetic resonance imaging scan even if their name has been removed from the record, thus enabling their re-identification from publicly available photographs[21]. Figuratively, this is analogous to considering passport photos without additional information not as personal data. Arguably, this highlights the tension between what is considered 'private' in a legal sense and what individuals consider acceptable in terms of informational self-determination. We thus contend that AI systems that process sensitive data should not only rely on de-identification techniques but also implement privacy-enhancing technologies (PETs), that is, technologies that furnish an objective or formal guarantee of privacy protection.

DP as the optimal privacy preservation

Among PETs, differential privacy (DP)[22] is considered the optimal protection for training AI models while moderating the privacy risk faced by participating patients due to its appealing properties: it provides a formal upper bound on the success of reconstructing data[23,24] and satisfies requirements imposed by regulations such as the General Data Protection Regulation concerning re-identification[19,25]. Moreover, the privacy guarantees of DP cannot be degraded through the use of side information or through post-processing (two notable vulnerabilities of traditional de-identification schemes). Last but not least, DP satisfies composability, that is, its guarantee degrades predictably when multiple DP algorithms are executed on the same dataset. This enables the concept of a 'privacy budget', which makes the cumulative re-identification risk quantifiable and can be set depending on policy or preference. We note that this ability to moderate risks stemming from AI applications is particularly beneficial, as it is also mandated by recent legal frameworks such as the European AI act[26]. These properties are leading to DP's increasing adoption in industry and government applications[27,28].

We remark that for a holistic workflow, additional PETs are advisable. Cryptographic techniques such as homomorphic encryption or secure multi-party computation can allow performing computations on data while ascertaining that only authorized instances can read the private information. However, these techniques are 'binary', that is, information is perfectly private (encrypted) or non-private (decrypted). In particular, at the latest at inference time, the information must be decrypted to be useful. In contrast, DP limits the probability that the output (gradient) can be correctly assigned to the input (data), which allows useful outputs at a guaranteed (but not perfect) level of privacy. Arguably, the most famous PET is federated learning, which provides a means to preserve data governance. However, without further protective measures, in particular DP, data can be reconstructed, and thus data governance is again not maintained. An overview can be found in ref. 29.

Despite these benefits, the effective and efficient implementation of DP in large-scale AI systems also presents a series of challenges. DP has been criticized for the fact that the choice of an appropriate privacy budget is delicate. Higher budgets correspond to less privacy protection and thus an increased risk of successful attacks, while lower budgets limit the information available for training. This introduces new challenges, namely a trade-off between privacy and model performance, that is diagnostic accuracy for a given use case. Furthermore, this trade-off also depends on the specific input data and learning task, which can vary drastically between scenarios. Arguably, concerns about reduced model performance are a probable reason why, despite its benefits, DP is not yet widely implemented in medical AI. After all, finding a trade-off between diagnostic accuracy and privacy represents a complex technical and ethical dilemma. This dilemma is best understood as DP is underlain by a worst-case set of assumptions. These assumptions, also called a threat model, include an adversary who is able to deeply manipulate and interfere with the dataset, the training process, model architecture and (hyper-)parameters, and has access to all parameters of the DP algorithm (mechanism). Moreover, the canonical DP adversary is not assumed to execute a data reconstruction attack but a much simpler type of attack, namely a membership inference attack, which attempts to determine whether a specific individual's data (which is available to the adversary) was included in the training dataset or not. Since there are only two possible outcomes of such an attack (member/non-member), membership inference must only reveal a single bit of information compared with a data reconstruction attack, which must successfully reveal a much larger record (for example, an image). Although worst-case assumptions are prudent for the theoretical modelling of adversaries, the DP threat model is unlikely to ever be encountered in practice. Moreover, the aforementioned membership inference attack in which the adversary has access to a target record and tries to determine whether it was used for training a specific model is arguably of very low practical relevance. Instead, data reconstruction attacks are probably perceived as a substantially more relevant privacy threat by patients. Moreover, realistic adversaries in the medical setting (where data is strongly guarded) can probably be assumed to not have access to the training data (as they would have little incentive to attack a model otherwise).

Table 1 | Overview of the capabilities of an adversary in the threat models analysed in this study

	Worst case	Relaxed	Realistic
Model architecture and weight	Yes	Yes	Yes
Hyper-parameter	Yes	Yes	Yes
Dataset access	Yes	Partially	No
Perfect reconstruction algorithm	Not applicable	Yes	No
Risk analysis	Theoretical	Theoretical	Empirical

In this Article, we investigate whether the aforementioned typical DP threat model might be too pessimistic for practical use cases and thus impose unnecessary privacy/performance trade-offs. To investigate this hypothesis, we study the privacy/performance characteristics of AI models trained on large-scale medical imaging datasets under more realistic threat models that still allow for strong privacy protection but represent a 'step down' from the worst-case assumptions of DP. Our main finding is that, even in complex medical imaging tasks, it is possible to train AI models with excellent diagnostic performance while still defending against data reconstruction attacks and thus a likely patient re-identification. We achieve this by training models under privacy budgets that would be considered too large to offer any protection against the threats considered under the worst-case DP threat model. This supports a recommendation for training AI models with DP protection by default. Therefore, although more restrictive privacy budgets than the ones used in our study remain relevant for use cases in which protection against membership inference is explicitly required, there exists an additional option: when high model performance is required but cannot be achieved without relinquishing membership inference protection, our findings offer a compromise whereby an important and relevant class of attacks can be defended against while fulfilling the requirement for high diagnostic accuracy.

As stated above, DP allows for a quantifiable reduction in the risk of privacy attacks associated with the training of AI models. In this work, we differentiate between three threat models, which we term worst case, relaxed, and realistic. DP, reconstruction risks and all threat models are described in detail in Supplementary Material A. An overview can be found in Table 1.

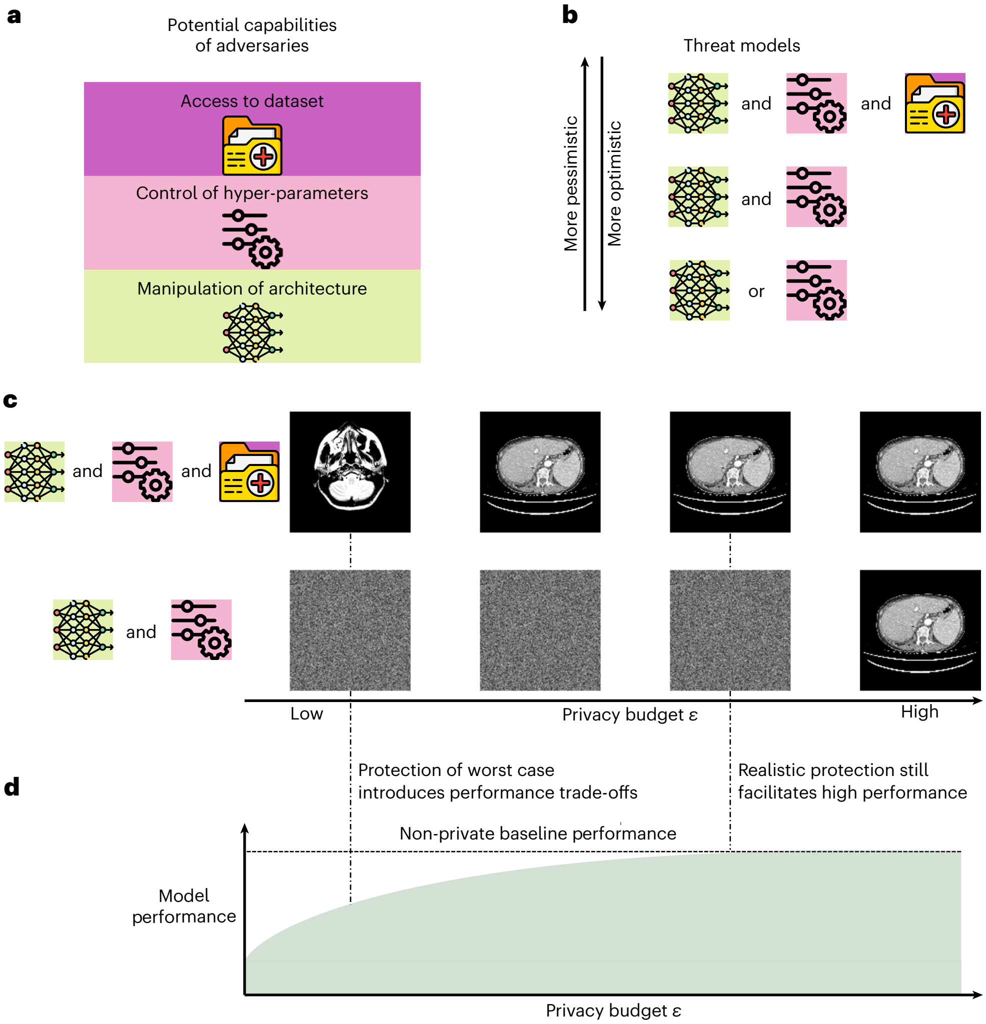
a
Potential capabilities of adversaries
Access to dataset
Control of hyper-parameters
Manipulation of architecture
b
Threat models
More pessimistic
More optimistic
and
and
and
or
c
and
and
and
Low
Privacy budget ε
High
Protection of worst case introduces performance trade-offs
Realistic protection still facilitates high performance
d
Non-private baseline performance
Model performance
Privacy budget ε

Fig. 1 | Comparison of a worst-case and a realistic threat model. a, Adversaries can have various capabilities depending on the setting. **b**, The combination of the adversary's capabilities defines the threat model. In a worst-case analysis, they have all capabilities. However, access to the database is a pessimistic, practically irrelevant scenario. **c**, The necessary privacy protection depends on the threat model. In a worst-case threat model, the adversary only needs to match the model and gradient to an image in the database. In a practically more relevant scenario, the image must be reconstructed from the model and gradient. Here, much less privacy protection is necessary. **d**, The more stringent the privacy protection is chosen, the higher the impacts on the model performance are. Thus, if a realistic threat model is considered appropriate, models can perform better.

The key contribution of our work is to investigate the realistic risks posed by a type of adversary who is still very powerful but can be reasonably assumed to exist in real-world medical AI model training use cases. An overview is displayed in Fig. 1. In the next section, we will show that perfectly defending against such adversaries is possible while maintaining a diagnostic model performance competitive with that of a model trained without any privacy protection.

Results

Set-up

Our evaluation focuses on how various privacy risks on multiple real-world characteristic datasets (compare Table 2) correlate with the algorithm's performance. We provide details on the datasets and our rationale for choosing these in Supplementary Material B1 and on the evaluation metrics in B2. First, we show the correlation of the AI performance on our datasets with privacy budgets. Second, we illustrate the implications of a certain privacy budget in a risk profile, summarizing the reconstruction risk under different threat models. We recall that a threat model corresponds to the set of assumptions over the attacker, where we give the theoretical bounds for a worst-case and a slightly relaxed adversary. Both are more pessimistic than any real-world scenario. Thus, we add a third threat model representing the worst 'realistic' case.

Table 2 | Overview of characteristics of our datasets

Dataset	Task	Small	Imbalanced	Multi-modal
RadImageNet	Classification		✓	✓
HAM10000	Classification	✓	✓	
MSD Liver	Segmentation	✓	✓	

In Table 3, we list the best possible AI model performance and corresponding reconstruction risk for all datasets and privacy budgets. The risk is three-tiered: (1) The upper bound of a worst-case adversary. This is the maximum risk under this setting and cannot be increased by post-processing or side information. (2) The upper bound of a minimally relaxed adversary as introduced in ref. 24. (3) The reconstruction success of the real-world adversary. We argue that—for practical use cases—protection against such a real-world attacker suffices. By listing

https://doi.org/10.1038/s42256-024-00858-y

Table 3 | Comparison of performance to privacy risk over multiple datasets and privacy budgets

Privacy budget	Noise	Test MCC	Reconstruction risk		
ε at δ=8.0×10^{-7}	σ	Mean±s.d.	Worst case	Relaxed	Realistic
RadImageNet					
1	0.67	64.95±0.13%	0.00%	0.00%	0%
8	0.34	68.75±0.13%	0.04%	0.01%	0%
32	0.267	69.99±0.25%	13.18%	3.96%	0%
10^{12}	0.054	70.83±0.19%	100%	100%	0%
Non-private	0	71.83±1.86%	100%	100%	100%
HAM10000					
1	0.92	15.60±4.13%	0.03%	0.01%	0%
8	0.47	37.48±3.45%	1.22%	0.04%	0%
20	0.40	42.83±2.37%	22.30%	0.78%	0%
10^{9}	0.02	51.98±2.52%	100%	100%	0%
Non-private	0	51.66±1.38%	100%	100%	100%

		MSD Liver				
		Dice score liver	Dice score tumour	Reconstruction risk		
		Mean±s.d.	Mean±s.d.	Worst case	Relaxed	Realistic
1	9.97	42.84±1.83%	0.96±0.37%	1.66%	0.97%	0%
8	1.66	74.71±3.14%	3.01±0.96%	17.96%	3.68%	0%
20	0.96	79.06±2.17%	5.55±0.72%	74.24%	27.37%	0%
10^{9}	0.0054	91.20±0.23%	29.73±2.89%	100%	100%	0%
Non-private	0	91.58±0.41%	28.38±2.29%	100%	100%	100%

Test MCC denotes Matthew's correlation coefficient on the test dataset. For all performance metrics, we give the mean±s.d. over five runs with different random seeds. Reconstruction risk denotes the upper bounds for the risk of a successful reconstruction attack of a worst-case and minimally relaxed adversary, as well as the empirical success of one of the strongest 'realistic' attacks. An image is considered successfully reconstructed if the SSIM to any reconstruction is higher than 80%. Note that the noise multiplier σ is given for the empirical attack scenario where an adversary manipulated hyper-parameters in their favour. Noise multipliers for performance analysis are generally higher.

all three, we provide an overview of how the risk varies by changing assumptions about the adversary.

Performance trade-offs under varying privacy levels

Impacts on performance is substantial for small datasets. At first, we analyse the impact of a very restrictive (small) privacy budget of $\varepsilon = 1$ on the predictive AI performance on our datasets (Table 3). Across the board, we see that at these budgets, the impacts on the model performance are strong. Concretely, we find that on RadImageNet, a standard non-private AI model reaches 71.83% on average, while trained at such restrictive privacy guarantee we find an average Matthews' correlation coefficient (MCC) of 64.95%, which is still 90% of the non-private MCC score. The gap becomes much larger on the HAM10000 dataset, where the model performance, when trained with a very low privacy budget of $\varepsilon = 1$ is closely above the chance level at an MCC of 15.60%. Similarly, on the Medical Segmentation Decathlon (MSD) Liver dataset at restrictive privacy budgets, the average Dice score for the liver drops to 42.84% (non-private: 91.58%) and completely fails for the tumour with a Dice of 0.96%. This exemplifies the challenges of furnishing strong privacy protection when training AI models on small or difficult datasets.

Prediction quality under medium budgets depends on dataset. Next, we consider medium privacy budgets ranging from $\varepsilon = 8$ to $\varepsilon = 32$, which are typical choices in literature[30,31]. As ε is an exponential parameter (e^{ε}), larger values correspond to exponentially decreased privacy guarantees. For this reason, some argue that the guarantees provided by such medium budgets are meaningless[22,32].

At these privacy budgets, although the performance substantially increases compared with the extremely restrictive privacy budget, the private AI models never exactly match the non-private performance. On RadImageNet, the achieved result closely approaches the non-private baseline: at a privacy budget of $\varepsilon = 32$, the MCC is 69.99% versus 71.83% in the non-private case. Also, for HAM10000, performance is strongly improved at 42.83% MCC, yet still decreased by 9% compared with the non-private result. Lastly, in MSD Liver, the liver as a larger organ can now be learned up to a reasonable Dice score of 79.06% at $\varepsilon = 20$. However, it remains far from the non-private performance. The prediction quality of the tumour, which is a much smaller and more complex structure, is especially concerning. This leads to a poor segmentation quality and only achieves an average Dice score of 5.55%, which is unsuitable for real-world applications. Again, we note that performance trade-offs especially impact smaller and imbalanced datasets.

Performance trade-offs vanish under large privacy budgets. For very large privacy budgets, we observe that the gap between private and non-private performance disappears. We recall that HAM10000 and MSD Liver as small datasets are extremely challenging under restrictive DP conditions. When increasing the privacy budget to $\varepsilon = 10^9$, no statistically significant difference to the non-private model can be detected (P values: HAM10000: 0.36; and MSD Liver dataset liver: 0.10 and tumour: 0.29, Student's t-test). Only on RadImageNet, although the non-private model is still statistically significantly superior (P value: 0.001), the private model at an $\varepsilon = 10^{12}$ achieves 99% of the non-private baseline performance.

It is unsurprising that increasing the privacy budget mitigates the negative implications on the model performance. Hence, the question that must be asked is what level of privacy is necessary for a specific setting. This cannot be answered generally and must be carefully

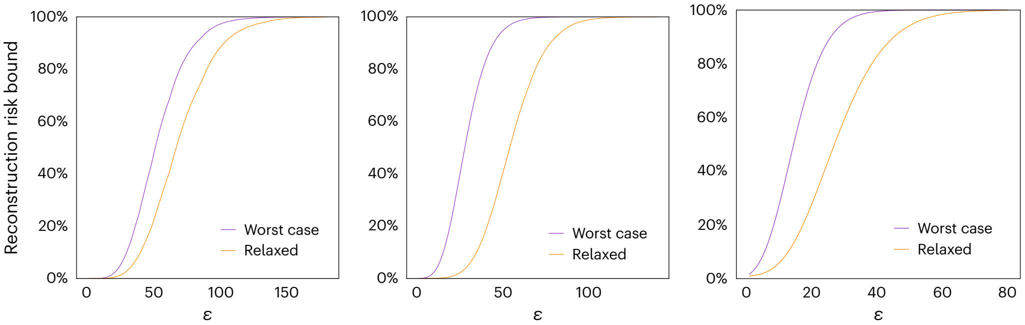

Fig. 2 | Theoretical reconstruction bounds for a worst-case and slightly relaxed adversary. From left to right: RadImageNet, HAM10000 and MSD Liver. We see that the mathematical upper bound for a reconstruction risk of a minimally relaxed threat model (orange) is already substantially lower compared with a worst-case setting (purple).

considered for each use case. Important for these considerations is which risks are associated with a certain privacy budget, which we analyse next.

Worst-case bounds require small privacy budgets

Although too pessimistic for most use cases, worst-case analyses have the advantage of a formal guarantee, that is, an absolute upper bound on the risk in this scenario. When analysing the theoretical worst-case (highest) success of reconstruction attackers, we find that for the large RadImageNet dataset for budgets $\varepsilon \leq 8$, the risk is <0.05%. However, already at $\varepsilon = 32$, the theoretical probability of the original data being reconstructed is 15%. Here, the smaller datasets are again at higher risk. While at $\varepsilon = 1$ the risk remains low, it strongly increases at $\varepsilon = 8$ for HAM10000 (0.03% to 1.22%) and MSD Liver (1.66% to 17.96%). At $\varepsilon = 20$ theoretically, up to 74.24% of all data samples of the MSD Liver dataset can be reconstructed.

However, even minimally relaxing the threat model assumptions decreases the risk associated with these privacy budgets drastically. We recall that under this relaxed threat model, the only change compared with the worst case is that the attacker does not know the sample that is reconstructed beforehand. Yet, for theoretical analysis, there is still the assumption that the reconstruction algorithm is either perfect or fails and the risk which is then calculated is the maximum rate where the attacker correctly decides if the reconstruction they obtained was indeed the dataset sample in question. This threat model is still too pessimistic for any real-world use case and the analysis is mostly for theoretical purposes. Still, such a minimal relaxation already gives a much more favourable risk profile, especially for medium privacy budgets. Exemplarily, the risk associated with $\varepsilon = 20$ diminishes from over 20% to less than 1% for the HAM10000 dataset. Similarly, the risk for the MSD dataset at $\varepsilon = 8$ decreases from 18% to 4%. A visualization of the risk difference in worst-case and relaxed threat models can be found in Fig. 2.

Empirical protection even at large privacy budgets

The previously discussed theoretical analyses show rapidly growing risks associated with small and medium privacy budgets. However, as discussed before, we argue that these analyses are too strict for any 'realistic' use case. Hence, we ask what the worst case of any practical scenario is and determine it to be a federated learning set-up, where a central server coordinates the learning on the data of distributed clients, which follow each training command sent by the server. This implies that the server can freely choose any network architecture and hyper-parameters. Note that any client who performs a simple check would notice such a malicious server. For such cases, attacks have been shown in literature, which analytically can recover the model input perfectly[8,9]. Moreover, it has been shown that these attacks can be transferred to corrupted pre-trained models[17]. We employ these attacks as empirical risk assessments. To measure the reconstruction success, we use the structural similarity (SSIM) score, which is a standard metric for image similarity[33].

In contrast to the aforementioned theoretical risk bounds, we find that, for practical attacks, even privacy budgets considered meaningless ($\varepsilon > 10^9$) can provide effective protection against reconstruction. In Fig. 3, left, we plot how many dataset images are below an increasing SSIM error per privacy budget. It can be thought of as the cumulative distribution function of reconstruction errors. We observe that, for all datasets without the addition of DP constraints, nearly all images can be reconstructed perfectly. As soon as some privacy guarantee is introduced, even very generous budgets at an $\varepsilon \approx 10^9$ provide empirical protection against the reconstruction of data samples. Furthermore, confirming previous works[8,34], our threat model is still extremely powerful. A server without the control of hyper-parameters but still over the model architecture already imposes a substantially lower reconstruction risk. If the server does not set the batch size to one but is set to the real training batch size, for example, on the RadImagenet dataset even in the non-private case we could only reconstruct less than 5% of all images at a batch size of 3,328. We note that such large privacy budgets, which are near-universally shunned as being meaningless, still offer empirical protection. In other words, even a 'pinch of privacy' has drastic effects in practical scenarios. Complemented by the finding that performance trade-offs nearly disappear in these settings, this signifies a potential compromise between protection and usability.

Discussion

In this study, we explore the relationship between privacy risks and AI performance in sensitive applications such as medical imaging. Currently, practitioners are confronted with trade-offs between AI performance, privacy protection and computational efficiency, where no solution has so far been able to accomplish all of these goals. Previous work showed that DP training profits much more than standard AI training from a higher number of training steps[30]. By increasing privacy budgets, practitioners can reach similar trade-offs with fewer training steps, which further allows a broader use for practitioners without substantial compute resources. Moreover, prior work also showed that pre-training on a 4 billion image dataset allows models to transfer to private datasets[35]. However, in practice this is typically infeasible due to limited access to such large datasets or the computational resources to train such a model. Furthermore, such data scales only exist for natural two-dimensional images but not yet for three-dimensional images, which are typical in medical imaging. Therefore, often the choice remains for practitioners to prioritise privacy and sacrifice performance or to put sensitive data at risk of being leaked. Currently, there is no clear method to balance these two objectives, leaving practitioners without guidance. To make informed decisions on these trade-offs, broad discourse involving ethicists, lawmakers

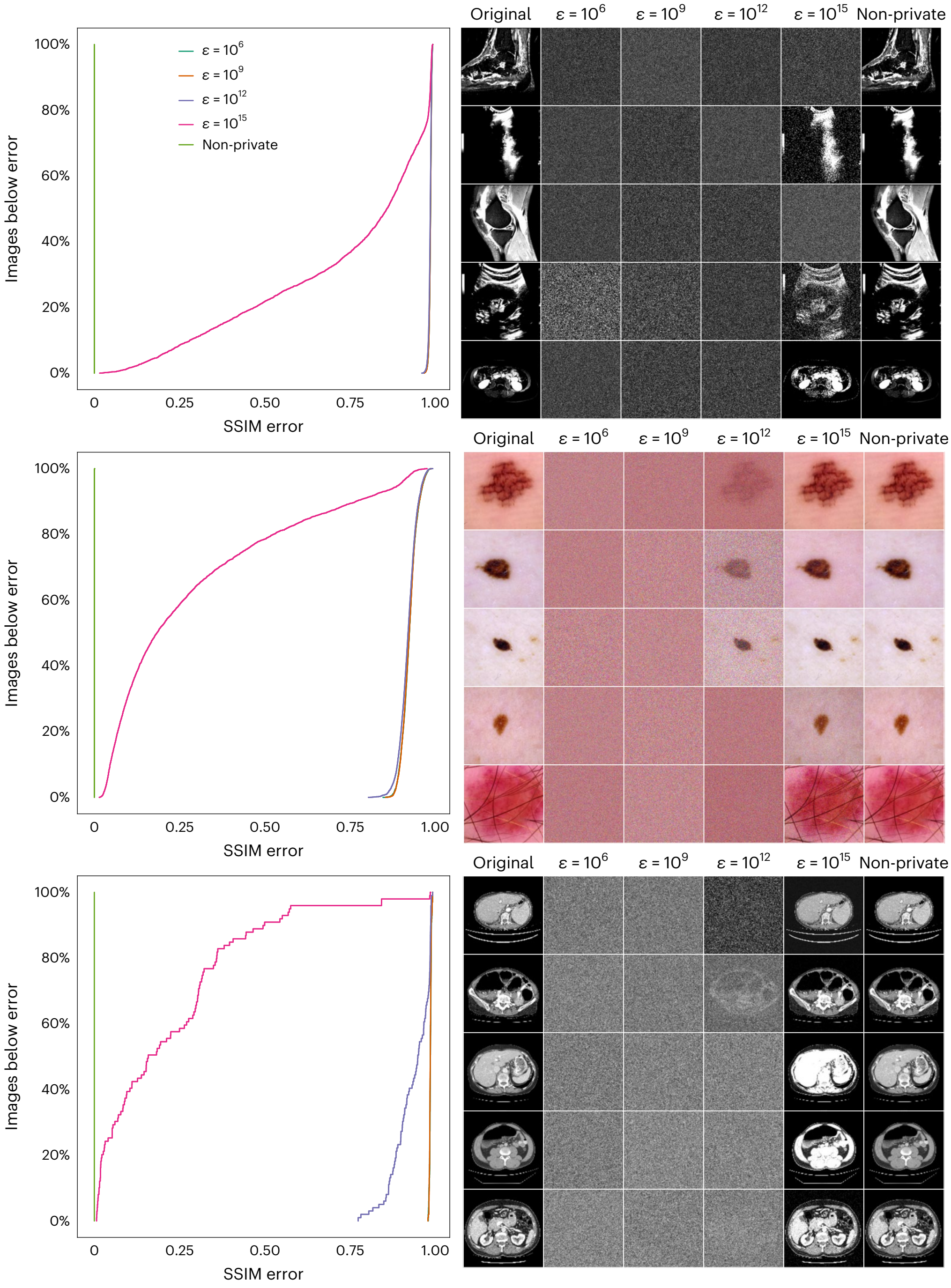

Fig. 3 | Reconstruction threat analysis for three datasets. Each row shows one dataset. From top to bottom: RadImageNet, HAM10000 and MSD Liver. Left: the cumulative number of images that have, in an empirical reconstruction, a SSIM difference lower than the value on the *x* axis. Note that it is the SSIM reconstruction error and thus perfect at 0 and worst at 1. Exemplarily, we see that on the MSD dataset at a reconstruction error of 10% all non-private (green) images, 39% at $\varepsilon = 10^{15}$ (pink) and none at more restrictive privacy guarantees can be reconstructed Right: the top five images with the best reconstruction score and their corresponding best reconstruction at various privacy budgets.

and the general population is crucial. A prerequisite of this dialogue is understanding the risks associated with specific privacy budgets and the potential trade-offs in AI performance. Our study across three representative medical imaging datasets lays the foundation for this conversation. We find that real-world data reconstruction risks can be averted without performance trade-offs. In fact, privacy–performance trade-offs have so far always been based on worst-case assumptions, which do not overlap with realistic training settings. We postulate that it is more critical to prevent data reconstruction in real-world settings, and show that for workflow de-risking, large privacy budgets suffice. Even more, we find that the trade-off between privacy risks and model performance vanishes when using such large but protective privacy budgets.

It is known from previous works[23,36–38] that PETs formally protect AI models in sensitive contexts from reconstruction attacks. While we note that our results are empirical, it is apparent that DP training with minimal guarantees still provides better protection than non-private training. Considering this finding, it seems negligent to train AI models without any form of formal privacy guarantee. We note that the threat model we consider is probably still stronger than attackers encountered in practical attack scenarios. In a slightly different threat model, where an adversary only has black-box access to the final trained weights of a model but has an image prior containing the true target point, ref. 23 found that large privacy budgets in the order of the dimensionality of the data suffice to prevent reconstruction attacks. Similarly, ref. 32 found that against reconstruction attacks, noise multipliers which otherwise would be seen as vacuous, suffice. Furthermore, ref. 39 studied the reconstruction of discrete data and found that privacy budgets can be much larger than previously thought to effectively defend against reconstruction attacks. However, for our threat model, we find even much larger privacy budgets than the aforementioned to suffice and, without a theoretical lower bound, the possibility exists that future attacks could achieve success closer to the upper bound. Owing to this, we explicitly warn readers to take our results as a carte blanche to use arbitrarily high privacy budgets. The truth lies in the middle: if the alternative is to not use any privacy at all, rather use DP with a very high budget.

We remark that the effectiveness of the DP protection against attacks at a fixed clipping norm, batch size, training duration and training set size depends only on the noise multiplier. This is a consequence of how DP budgets are accounted. For example, in the Rényi-DP (RDP) accountant[40] used in our work, one step is $(\alpha, q^2\frac{2\alpha C^2}{\sigma^2})$-RDP for appropriate values of the parameters α the order of the Rényi divergence, q, the subsampling rate (that is, batch size divided by training set size), C, the clip gradient norm and σ, the noise multiplier. However, our empirical results suggest that for all other factors being constant, even small noise multipliers, which imply very large privacy budgets, are sufficient to protect against reconstruction attacks and facilitate high-performing AI models. We also observed that the AI performance loss introduced by DP tends to be smaller on larger datasets due to less injected noise per sample and more information to achieve a certain privacy budget at consistent hyper-parameters. Yet, many medical datasets are inherently small. This can have negative consequences for the applicability of such networks in clinical practice. For models to be effectively trained on such challenging datasets, when pre-training is not possible for reasons of data availability or computational resources, our techniques reach a limit indicating a potential need to either accept elevated privacy risks or obtain access to more data. The solution to both problems might go hand in hand with more robust mathematical guarantees safeguarding data privacy. In such a scenario, we anticipate that patients may be more inclined to share their data, thereby allowing large-scale medical AI training. In such a scenario, the privacy–performance trade-offs presented might even be more favourable than our findings indicate. This would be complemented by a workflow where multiple PETs are employed to enable various aspects to privacy. For example, a system using federated learning to assert the data governance remains at the original hospital, secure aggregation to conceal contributions from different sites and DP to limit the private information of single patients demonstrated in previous works[36] would provide a holistic workflow.

We note that our choice of datasets and architectures is motivated by medical imaging settings. In those settings, typically computational resources are limited and data are scarce. In fact, we are convinced that the widespread use of such methods will only ensue once they can be used by the majority of practitioners who typically lack access to large computing clusters. Hence, we carefully designed our study to cover typical and representative medical problems to provide a holistic analysis with trade-offs in computational resources. Under these considerations, we limited ourselves to a few model architectures that are known to be trained efficiently (ResNet, DenseNet and U-Net) and datasets that represent a broad range of typical problems.

An additional technical limitation stems from the fact that the authors of the RadImagenet dataset[41] mention that some patients contributed multiple images. However, we have no information about image-to-patient correspondence. As we calculate the privacy guarantees over the dataset per image, the per-patient privacy guarantee depends on the number of images one patient contributed and might be lower.

In conclusion, we show that even the use of nominally loose privacy guarantees still provides substantially better protection than standard AI training, while achieving comparable performance. This can facilitate a compromise between provable risk management and performance trade-offs, which previously prevented the breakthrough of DP. Further research should be directed towards analysing various threat models beyond the worst case. Only by illuminating the risks of multiple scenarios, the basis for a broad discussion among ethicists, policymakers, patients and other stakeholders is provided regarding how to trade-off privacy and performance as fundamental goals of AI in sensitive applications.

Methods

In this section, we report all the details necessary for our experiments on training models in a differentially private way on our datasets as well as the procedures to analyse risk profiles. Furthermore, we describe the rationale for several choices in our study design and describe hyper-parameters necessary for reproducibility.

Data

In Supplementary Material A, we describe characteristics of typical medical datasets. We note, that these characteristics partially amplify the negative performance impact by the constraints introduced by DP. Broadly speaking, at a constant clipping norm the amount of introduced noise during the DP process determines the negative impact on the AI performance. At any privacy budget, the injected noise increases if more training steps are performed or if a higher sampling rate, that is, the ratio between batch size and dataset size, is used. However, the batch size is typically irrespective of the dataset size, which implies that smaller datasets typically have higher sampling rates. Furthermore, they often require more training epochs, that is, the amount of times the entire dataset was (on average) presented to the network. As a consequence, the amount of noise that is injected when training on small datasets compared with larger ones is increased and higher performance penalties are expected. Furthermore, DP bounds the magnitude any single sample on the training. This is important for training with imbalanced datasets with underrepresented classes, which often suffer an additional performance loss[42].

For detailed descriptions of the datasets we refer to the original publications[41,43–45]. In the following, we describe modifications we performed and the effects on the data distribution.

For the HAM10000 dataset[43], we merged classes into whether there is indication for immediate treatment, which is still a medically important distinction. By this we convert the multi-class classification problem into a highly imbalanced binary classification problem. We categorized them here as follows:

Treatment indication	
Immediate	**Not immediate**
Actinic keratoses and intra-epithelial carcinomas	Melanocytic nevi
Basal cell carcinomas	Benign keratinocytic lesions
Melanomas	Dermatofibromas
	Vascular lesions

In total, this dataset has 10,015 images, of which 1,954 are labelled for immediate treatment and 8,061 are not.

Model training

All of our experiments were performed using an NAdam optimizer, which is extremely robust to learning rate changes allowing us to keep a consistent learning rate of 2e^{-3}. Input data were always normalized with the mean and standard deviation of all images in the training set. For each dataset, we perform a hyper-parameter search, where we evaluate for one privacy level ($\varepsilon = 8$) and the non-private training the optimal setting for architecture, batch size, loss weighting and augmentation. In the non-private case, we perform an early stopping strategy to determine the number of epochs. In the private case, this is not possible as the number of epochs directly influences the amount of added noise. However, previous works showed that longer training almost always yields better results[30]. Yet, to limit training time, we also search for the point of saturation. Also for reasons of computational complexity, we assume that the optimal settings for these parameters transfer to all other privacy regimes. Furthermore, we limit the choice of architectures to a ResNet-9 with ScaleNorm and a WideResNet40-4, which have in previous literature been proven to be especially suited for differentially private training[30,46]. In the segmentation case, we limit ourselves to a standard U-Net[47,48], where we optimize the number of channels on the bottleneck. We then evaluate for each privacy setting separately the optimal clipping norm. Again for reasons of computational complexity, we evaluate this after one epoch and assume it transfers to longer trainings. Finally, we train for each setting five models with different random seeds and report the mean and standard deviation of the respective performance metric.

All our models are trained from 'scratch', that is, we have not pre-trained on any other dataset. This is because there is no 'good choice' of a dataset for pre-training. ImageNet, which for most computer vision tasks is the standard, is not very effective for medical imaging tasks[41]. Large public databases for pre-training are scarce and only available for a few tasks. Furthermore, pre-training on non-public medical databases is unacceptable, as it risks leaking the information from the pre-training data, which could be just as private[49,50].

We used the Opacus[51] library for accounting the privacy loss. In particular, we used an RDP accountant, as it provides numerically the most stable implementation. We used an extension of the objax library[52] as implementation for the DP-Stochastic Gradient Descent algorithm.

We open source the program code used for this paper at https://github.com/a1302z/RePrAAIMI.

RadImagenet. As described in the 'Model training' section, we analysed the architecture, number of epochs, batch size, loss and multiplicity for the non-private and one private setting ($\varepsilon = 8$). For the non-private case, we found a WideResNet40-4 using an unweighted loss function, a batch size of 16 and random vertical (probability of augmentation (P_{aug}) = 0.2) and horizontal flips ($P_{\mathrm{aug}} = 0.1$) as augmentation to yield the best results. To determine the number of epochs, we used an early stopping strategy with a patience of five epochs and 0.1% improvement threshold. For the private case, a ResNet-9 trained for 50 epochs, using an unweighted loss function, using an augmentation multiplicity of four again with random vertical ($P_{\mathrm{aug}} = 0.2$) and horizontal ($P_{\mathrm{aug}} = 0.2$) flips with a batch size of 3,328 yielded best results. The clipping norm was tuned for each budget separately and was set as follows:

ε	1	8	32	1e^{12}
Clip norm	6.46	5.66	5	3.75

HAM10000. For the modified HAM10000 dataset, we found the ResNet-9 to perform best in private and non-private settings. In the non-private case, we trained with a weighted loss function at a batch size of 32 using random vertical flips ($P_{\mathrm{aug}} = 0.5$) as augmentation. We trained using an early stopping strategy using a patience of 50 epochs at a minimal improvement threshold of 0.1%. For the private case, we used an unweighted loss function at a batch size of 2,048 and trained for 100 epochs. We used the same augmentations as in the non-private case for a privacy level of $\varepsilon = 10^9$, for all others, we did not use augmentations. Clipping norms are as follows:

ε	1	8	20	1e^{9}
Clip norm	18	8.5	9.5	9

MSD Liver. For the MSD Liver dataset, we found for both private and non-private cases a U-Net with 16 channels and no augmentations to perform best. In the non-private case we used a weighted loss function (background: 0.1; liver: 0.4; tumour: 0.5) and trained at a batch size of two. Again, we employed an early stopping strategy with a patience of 50 epochs and a minimal improvement threshold of 0.1%. In the private case, we trained at a batch size of one for 500 epochs. For privacy budgets $\varepsilon \leq 20$ we used an unweighted loss function, for higher privacy budgets we used the same weighting as in the non-private case.

ε	1	8	20	1e^{9}
Clip norm	0.0004	0.046	0.0015	0.33

Reconstruction risk analysis

In our empirical reconstruction attacks, there is no clear way to evaluate whether a specific sample was reconstructed. For each input batch consisting of N samples, we receive M reconstructions. We evaluate this by calculating the pairwise distance between all data samples and reconstructions and assigning each input the reconstruction with the lowest distance. However, this approach loses meaning in the case of images, which have no structure but are entirely dark. This is the case for the RadImagenet dataset, where we put a constraint that only data samples are considered that contain more than 10% non-zero pixels.

We evaluate the practical reconstruction success by using a principle demonstrated in previous literature[8,9] adapted to our use case. The network architecture is slightly modified by prepending two linear layers in front of the actual network architecture. The first takes all input image pixels as input and projects them to an intermediate representation of N bins. In our experiments, we set $N = 10$. This intermediate representation is afterwards projected again to the number of all pixels and re-sized to the original image shape. To each of the outputs, the mean of the intermediate representations is added. Afterwards, it can be processed as usual by the remaining neural network. As our adversary is assumed to have control over all hyper-parameters, they can set the batch size to one and by that enforce that no reconstruction of two images overlap. If now a gradient is calculated over the network, which is non-zero for the weights W_i and biases b of the first linear layer, the input x can be analytically recovered by $x = \nabla_{W_i}\mathcal{L} \oslash \frac{\partial \mathcal{L}}{\partial b}$, where $\oslash$ is the element-wise division. We note that, for this attack, it is irrelevant what network architecture comes after this imprint block. We used implementations provided by ref. 53.

The reconstruction error, which we use as basis for the risk analysis in this paper, is the minimum reconstruction error between a data sample to any reconstruction that was derived from a gradient containing the data sample.

Choice of privacy budgets

For our experiments on the utility trade-off, we chose several privacy budgets. We note that this choice was arbitrary. For all experiments, we used a $\delta = 8 \times 10^{-7}$. For all settings, we evaluated $\varepsilon = 1$ and $\varepsilon = 8$, which are standard values in the literature[30,31,46]. Furthermore, we calculate the theoretical reconstruction bound of the worst case and relaxed threat models. As the already included privacy budgets at $\varepsilon = 1$ and $\varepsilon = 8$ already showcase very low reconstruction bounds, we add one more privacy level for all datasets, where a large amount of samples is already at risk of being reconstructed. In addition, we report a privacy budget $\varepsilon = 10^{3N}, N \in \mathbb{N}$, where the characteristic reconstruction robustness curve is still similar to random noise.

Environmental impact

Lastly, we would like to give a rough estimate of the climate impact of this study. We assume the average German power mix that as of 2021 according to the German Federal Environment Agency corresponds to 475 g CO_2e kWh^{-1} (ref. 54) Only the final RadImagenet trainings (no hyper-parameter optimization) ran on eight NVIDIA A40s, where we assume a power consumption of 250 W on average, each for almost 4 days, five privacy levels and five repetitions. Hence, this amounts to around 960 kWh and thus more than 450 kg of CO_2e. This almost equals a return flight from Munich to London. Hence, we tried to limit our hyper-parameter searches to the necessary. In total, we assume that this study produced at least 2 tons of CO_2e.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets used in this study are published and publicly available. Access to RadImageNet[41] must be requested at https://www.radimagenet.com/. The HAM10000 dataset[43] is available at https://doi.org/10.7910/DVN/DBW86T. The MSD Liver dataset[44,45] is available at http://medicaldecathlon.com/ and https://doi.org/10.1038/s41467-022-30695-9.

Code availability

Our program code is available at https://github.com/a1302z/RePrAAIMI and permanently archived under https://doi.org/10.5281/zenodo.11184978 ref. 55. Furthermore, we created a modified version of[53], which is available at https://github.com/a1302z/objaxbreaching and https://doi.org/10.5281/zenodo.11184998 ref. 56.

References

1. Lång, K. et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **24**, 936–944 (2023).
2. Wang, G. et al. Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. *Nat. Med.* **29**, 2007–2018 (2023).
3. Al-Zaiti, S. S. et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. *Nat. Med.* **29**, 1804–1813 (2023).
4. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
5. Yao Jiang, L. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
6. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. Inverting gradients—how easy is it to break privacy in federated learning? *Adv. Neural Inf. Process. Sys.* **33**, 16937–16947 (2020).
7. Yin, H. et al. See through gradients: image batch recovery via gradinversion. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16337–16346 (2021).
8. Fowl, L., Geiping, J., Czaja, W., Goldblum, M. & Goldstein, T. Robbing the fed: directly obtaining private data in federated learning with modified models. In *Tenth International Conference on Learning Representations* (2022).
9. Boenisch, F. et al. When the curious abandon honesty: federated learning is not private. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)* 175–199 (IEEE, 2023).
10. Wang, Kuan-Chieh et al. Variational model inversion attacks. *Adv. Neural Inf. Process. Syst.* **34**, 9706–9719 (2021).
11. Haim, N., Vardi, G., Yehudai, G., Shamir, O. & Irani, M. Reconstructing training data from trained neural networks. *Adv. Neural Inf. Process. Syst.* **35**, 22911–22924 (2022).
12. Carlini, N. et al. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)* 5253–5270 (2023).
13. Buzaglo, G. et al. Deconstructing data reconstruction: multiclass, weight decay and general losses. In *Thirty-Seventh Conference on Neural Information Processing Systems* (2023).
14. Hatamizadeh, A. et al. Do gradient inversion attacks make federated learning unsafe? *IEEE Trans. Med. Imaging* **42**, 2044–2056 (2023).
15. Chen, H., Zhu, T., Zhang, T., Zhou, W. & Yu, P. S. Privacy and fairness in federated learning: on the perspective of tradeoff. *ACM Comput. Surv.* **56**, 1–37 (2023).
16. Usynin, D., Rueckert, D. & Kaissis, G. Beyond gradients: exploiting adversarial priors in model inversion attacks. *ACM Trans. Priv. Secur.* **26**, 1–30 (2023).
17. Feng, S.& Tramèr, F. Privacy backdoors: stealing data with corrupted pretrained models. In *International Conference on Machine Learning (ICML)* (2024).
18. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* 111–125 (IEEE, 2008).
19. Cohen, A. & Nissim, K. Towards formalizing the GDPR's notion of singling out. *Proc. Natl Acad. Sci. USA* **117**, 8344–8352 (2020).
20. Cohen, A. Attacks on deidentification's defenses. In *31st USENIX Security Symposium (USENIX Security 22)* 1469–1486, (2022).
21. Schwarz, C. G. et al. Identification of anonymous mri research participants with face-recognition software. *N. Engl. J. Med.* **381**, 1684–1686 (2019).
22. Dwork, C. et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2014).
23. Balle, B., Cherubin, G. & Hayes, J. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)* 1138–1156 (IEEE, 2022).
24. Kaissis, G., Hayes, J., Ziller, A. & Rueckert, D. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy. *CoRR* abs/2307.03928 (2023).
25. Nissim, K. Privacy: from database reconstruction to legal theorems. In *Proc. 40th ACM SIGMOD–SIGACT–SIGAI Symposium on Principles of Database Systems* 33–41 (2021).
26. *Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, document 52021PC0206* (European Parliament and of the Council, 2021).
27. Foote, A. D., Machanavajjhala, A. & McKinney, K. Releasing earnings distributions using differential privacy: disclosure avoidance system for post-secondary employment outcomes (PSEO). *J. Priv. Confidential.* **9**, 2 (2019).
28. Aktay, A. et al. Google COVID-19 community mobility reports: anonymization process description (version 1.1). Preprint at https://arxiv.org/abs/2004.04145 (2020).
29. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
30. De, S., Berrada, L., Hayes, J., Smith, S. L. & Balle, B. Unlocking high-accuracy differentially private image classification through scale. Preprint at https://arxiv.org/abs/2204.13650 (2022).
31. Sander, T., Stock, P. & Sablayrolles, A. Tan without a burn: scaling laws of dp-sgd. In *International Conference on Machine Learning* 29937–29949 (PMLR, 2023).
32. Stock, P., Shilov, I., Mironov, I. & Sablayrolles, A. Defending against reconstruction attacks with Rényi differential privacy. *CoRR* abs/2202.07623 (2022).
33. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).

34. Usynin, D., Rueckert, D., Passerat-Palmbach, J. & Kaissis, G. Zen and the art of model adaptation: low-utility-cost attack mitigations in collaborative machine learning. *Proc. Priv. Enhancing Technol*. **2022**, 274–290 (2022).
35. Berrada, L. et al. Unlocking accuracy and fairness in differentially private image classification. Preprint at https://arxiv.org/abs/2308.10888 (2023).
36. Kaissis, G. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach.Intell*. **3**, 473–484 (2021).
37. Ziegler, J., Pfitzner, B., Schulz, H., Saalbach, A. & Arnrich, B. Defending against reconstruction attacks through differentially private federated learning for classification of heterogeneous chest x-ray data. *Sensors* **22**, 5195 (2022).
38. Hayes, J., Mahloujifar, S. & Balle, B. Bounding training data reconstruction in DP-SGD. In Proc. *37th Conference on Neural Information Processing Systems* (OpenReview.net, 2023).
39. Guo, C., Sablayrolles, A. & Sanjabi, M. Analyzing privacy leakage in machine learning via multiple hypothesis testing: a lesson from fano. In *International Conference on Machine Learning* 11998–12011 (PMLR, 2023).
40. Mironov, I., Talwar, K. & Zhang, L. Rényi differential privacy of the sampled Gaussian mechanism. Preprint at https://arxiv.org/abs/1908.10530 (2019).
41. Mei, X. et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiol. Artif. Intell*. **4.5**, e210315 (2022).
42. Bagdasaryan, E., Poursaeed, O. & Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Adv. Neural Inf. Process. Syst*. **32**, (2019).
43. Tschandl, P., Rosendahl, C. & Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 1–9 (2018).
44. Simpson, A. L. et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. Preprint at https://arxiv.org/abs/1902.09063 (2019).
45. Antonelli, M. et al. The medical segmentation decathlon. *Nat. Commun*. **13**, 4128 (2022).
46. Klause, H., Ziller, A., Rueckert, D., Hammernik, K. & Kaissis, G. Differentially private training of residual networks with scale normalisation. In *Theory and Practice of Differential Privacy Workshop* (ICML, 2022).
47. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015*. Part III 18, 234–241 (Springer, 2015).
48. Çiçek, Özgün, Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D u-net: learning dense volumetric segmentation from sparse annotation. In *Proc. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016*. Part II 19, 424–432 (Springer, 2016).
49. Abascal, J., Wu, S., Oprea, A. & Ullman, J. Tmi! finetuned models spill secrets from pretraining. In *The Second Workshop on New Frontiers in Adversarial Machine Learning* (2023).
50. Tramèr, F., Kamath, G. & Carlini, N. Considerations for differentially private learning with large-scale public pretraining. Preprint at https://arxiv.org/abs/2212.06470 (2022).
51. Yousefpour, Ashkan, et al. Opacus: user-friendly differential privacy library in PyTorch. Preprint at https://arxiv.org/abs/2109.12298 (2021).
52. Objax. *Objax Developers* https://github.com/google/objax (2022).
53. Wen, Y., Geiping, J. & Fowl, L. Breaching. *GitHub* https://github.com/JonasGeiping/breaching (2023).
54. Icha, P., Lauf, T. & Kuhs, G. Entwicklung der spezifischen Treibhausgas-Emissionen des deutschen Strommix in den Jahren 1990–2021. *Umweltbundesamt Dessau-Roß*lau (2022).
55. Ziller, A., Kaissis, G. & Stieger, S. a1302z/repraaimi. *Zenodo* https://doi.org/10.5281/zenodo.11184978 (2024).
56. Ziller, A. objaxbreaching. *Zenodo* https://doi.org/10.5281/zenodo.11184998 (2024).

Acknowledgements

The authors acknowledge the following sources of funding for this work: A.Z. and G.K. are supported by the German Ministry of Education and Research (BMBF) under grant number 01ZZ2316C (PrivateAIM). T.T.M. and D.R. are supported supported by the ERC Grant Deep4MI (no. 884622). L.F.F. and R.B. are supported by the Federal Ministry of Education and Research (BMBF), Project 'NUM 2.0' (grant no. 01KX2121). R.B. is funded by the Federal Ministry of Education and Research (BMBF, grant no. 01ZZ2315B), Bavarian Cancer Research Center (BZKF, Lighthouse AI and Bioinformatics), German Cancer Consortium (DKTK, Joint Imaging Platform). G.K. and D.R. are funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts through the Munich Centre for Machine Learning. G.K. is supported by the Bavarian Collaborative Research Project PRIPREKI of the Free State of Bavaria Funding Programme 'Artificial Intelligence–Data Science' and from the German Academic Exchange Service (DAAD) under the Kondrad Zuse School of Excellence for Reliable AI (RelAI). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Furthermore, the authors thank J. Hayes for providing an implementation of the prior aware attack presented in ref. 38.

Author contributions

A.Z. conceptualized this study, wrote the program code, performed all experiments and prepared the paper. T.T.M. assisted in the preparation of the paper. S.S. assisted in the design of the program code. L.F.F. wrote program code for an efficient reconstruction matching and segmentation loss. J.B. helped to prepare the HAM10000 dataset for our purposes. R.B. and D.R. provided oversight. G.K. helped conceptualize this study and in the preparation of the paper, wrote code for the theoretical risk bounds and provided oversight. All authors revised the paper.

Funding

Open access funding provided by Technische Universität München.

Competing interests

The authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00858-y.

Correspondence and requests for materials should be addressed to Alexander Ziller.

Peer review information *Nature Machine Intelligence* thanks Holger Roth, Yiyu Shi, Tian Xia and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): Alexander Ziller, NATMACHINTELL-A231110141A

Last updated by author(s): May 14, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Permanently archived and cited computer code available at https://doi.org/10.5281/zenodo.11184978 and https://doi.org/10.5281/zenodo.11184998. Our code was developed in Python V3.10. We used Objax V1.6.0 for training our AI models. For accounting the privacy loss we used Opacus V1.4.0. Other packages that were used in this study are pytorch V2.0 scikit-learn V1.3.0, scikit-image V0.21.0, pandas V2.0.3, OpenCV V4.7.0, breaching V0.1.2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets used in this study are published and publicly available. Access to RadImageNet [28] must be requested at https://www.radimagenet.com/. The HAM10000 dataset [29] is available at https://doi.org/10.7910/DVN/DBW86T. The MSD Liver dataset [30,31] is available at http://medicaldecathlon.com/ and https://doi.org/10.1038/s41467-022-30695-9.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were predetermined by the public datasets used in this study. The datasets were chosen in order to represent typical medical AI workflows, with one multi-modal dataset, one small imbalanced dataset and a segmentation 3D dataset.
Data exclusions	None
Replication	Findings are deterministic with given data and seed for pseudo-random number generator.
Randomization	RadImagenet has a predetermined split into train, validation and test set, which ascertains that there is no data leakage with one patient in multiple cohorts. For HAM10000, we split randomly with a stratification by the class label in order to ensure that train, validation and test set approximately had the same distribution of classes. This was done using the train_test_split function of scikit-learn. For MSD 10000 we split randomly on a patient level.
Blinding	There are no groups which can be blinded. Therefore blinding is not applicable to this study.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For*

	studies involving existing datasets, please describe the dataset and source.
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*

Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
□	□	Public health
□	□	National security
□	□	Crops and/or livestock
□	□	Ecosystems
□	□	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
□	□	Demonstrate how to render a vaccine ineffective
□	□	Confer resistance to therapeutically useful antibiotics or antiviral agents
□	□	Enhance the virulence of a pathogen or render a nonpathogen virulent
□	□	Increase transmissibility of a pathogen
□	□	Alter the host range of a pathogen
□	□	Enable evasion of diagnostic/detection modalities
□	□	Enable the weaponization of a biological agent or toxin
□	□	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

□ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.
For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission
Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)
Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates
Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth
Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies
Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters
Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality
Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software
Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation
Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument
Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a | Involved in the study

□ □ Functional and/or effective connectivity

□ □ Graph analysis

□ □ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.