nature medicine

Matters arising

https://doi.org/10.1038/s41591-023-02439-9

Concerns about using a digital mask to safeguard patient privacy

Received: 18 November 2022

Accepted: 2 June 2023

Published online: 18 July 2023

Check for updates

Matthieu Meeus, Shubham Jain & Yves-Alexandre de Montjoye

ARISING FROM https://doi.org/10.1038/s41591-022-01966-1 (2022)

Sharing data is crucial for advancing medical research but should not come at the expense of patient privacy. Yang et al.[1] proposed to apply a digital mask (DM) to a facial image with the goal of retaining information relevant for medical diagnosis while 'irreversibly erasing identifiable features', making the data 'anonymous'[2]. The masking approach consists of a three-dimensional reconstruction from a two-dimensional facial image, to be rendered back as the DM. The paper shows that diagnosis of ocular conditions using masked reconstructions of facial videos is both accurate and consistent with the diagnosis on original (unmasked) videos. The authors show that the DM can evade AI-powered facial recognition systems, which underpins their claim that the method preserves privacy.

Although sharing data for medical diagnosis while preserving privacy is an important line of research, we believe the evaluation setup in Yang et al. to be inadequate, raising serious questions with regard to the risk to patient privacy posed by the proposed masking method. The facial recognition setup used by the authors as validation of the privacy-preserving capabilities of the DM assumes that an attacker attempting to identify a patient will try to match a mask to a database of faces (a Mask2Face approach) using a facial recognition algorithm. We argue that this setup and the corresponding empirical results reported by the authors do not properly evaluate the risk of reidentification. Indeed, a simple change to the setup, assuming the masking algorithm is available, allows an attacker to mask the faces before running a facial recognition algorithm on the now more comparable database of masked faces (a Mask2Mask approach).

The code made available by the authors is not sufficient to apply their masking technique to an image nor to evaluate the risk of reidentification. Similarly, the data they used to evaluate the preserving capabilities of their method are not available. To evaluate the risk of reidentification posed by the Mask2Mask approach, we instead used a similar linear face reconstruction model called FLAME[3], more specifically the RingNet implementation[4], to produce the facial masks. To evaluate the risk of reidentification, we used the Insightface implementation of the ArcFace[5,6] facial recognition model adopted by Yang et al. Finally, we used the YouTube Faces Database[7] as a dataset (Supplementary Information).

In this comparable setup, we first replicated the reidentification results obtained by Yang et al. We randomly sampled two frames from facial videos for each individual; then, we used one image in its original state as a reference image in the database, while the other image was used to compute the mask on a black background as the query image to be matched against the database (Mask2Face). Figure 1 shows that we obtained a rank-1 accuracy, the percentage of the time the algorithm identifies the right person in the database—the metric used by the authors for the risk of reidentification, of 0.7%, a value very similar to the 0.5% reported by Yang et al.

We then modified the setup to evaluate the risk posed by the Mask2Mask approach. In this setup, an attacker would obtain a rank-1 accuracy of 52% (Fig. 1) meaning that they can now correctly reidentify an individual more than half the time, an increase of 100-fold over the results reported by Yang et al. for the risk of reidentification (0.5%).

These results are furthermore only a lower bound on the actual risk. First, we used only the reconstructed face to reidentify patients in the protected database. The proposed method releases not only the reconstructed face but also the reconstructed eyeballs and eyelids. These are likely to provide further information to an attacker aiming to reidentify patients. Second, both our and the authors' reidentification results stem from readily available facial recognition algorithms. These are trained to identify individuals in pictures, based on detected facial patterns, but are not optimized for DM-reconstructed images. It is likely that better reidentification algorithms could be developed to reidentify masked patients[8,9]. An attacker leveraging the additional information available, such as eyeballs, and better reidentification algorithms is thus likely to be able to reidentify an individual with an even higher rank-1 accuracy than the one we report here.

Contrary to Yang et al.'s claims, our results show that the DM does not irreversibly erase identifiable features of a facial image. Anonymization requires, from both technical and legal perspectives, much more than an individual not being recognized by the human eye. Rather, GDPR Recital 26 (ref. 10) requires all means that are reasonably likely to be used by an attacker to be considered, and China's Personal Information Protection Law requires 'mak[ing] it impossible to distinguish specific natural persons and impossible to restore'[11]. Similarly, patients' privacy cannot, in general, be considered protected if it relies on an algorithm being kept secret now and forever[12]. In the case of the DM,

Department of Computing, Imperial College London, London, UK. e-mail: deMontjoye@imperial.ac.uk

https://doi.org/10.1038/s41591-023-02439-9

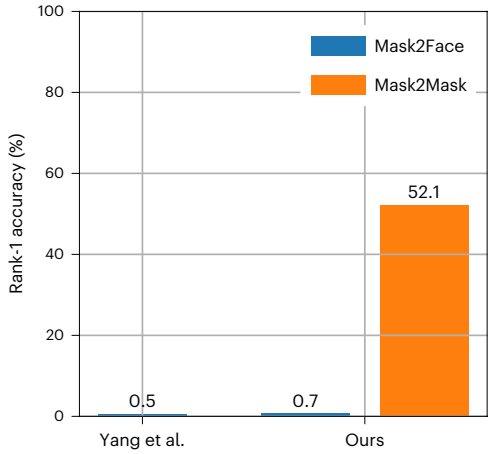

Fig. 1 | Mask2Mask achieves a reidentification accuracy of 52%. Rank-1 accuracy in facial recognition across 2 methods: (1) 'Mask2Face': mask as query image and original image as database image; (2) 'Mask2Mask': mask as query image and mask as database image. Results for Yang et al.[1] are based on analysis of 405 individuals, while results for our analysis ('ours') are based on analysis of 555 individuals.

the algorithm is published, relies on existing methods and is proposed to be deployed broadly.

Sharing data for research, in particular medical research, is highly beneficial to the scientific community and beyond, but cannot come at the expense of patient privacy and, ultimately, trust. While we appreciate the aims of Yang et al. to enable privacy-preserving patient diagnosis, ad hoc and inadequately tested methods have damaged patient trust before and put access to data for research at risk[13]. Although methods providing formal privacy guarantees are preferred, they are not always within reach or free from implementation issues. Any anonymization methods proposed therefore need to be extensively, and if possible adversarially, tested to ensure that privacy is preserved before data is shared.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-023-02439-9.

References

1. Yang, Y. et al. A digital mask to safeguard patient privacy. *Nat. Med.* **28**, 1883–1892 (2022).
2. Brierley, C. *'Digital mask' could protect patients' privacy in medical records.* https://www.cam.ac.uk/stories/digital-masks/ (University of Cambridge, 2022).
3. Li, T., Bolkart, T., Black, M. J., Li, H. & Romero, J. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* **36**, 194 (2017).
4. Sanyal, S., Bolkart, T., Feng, H. & Black, M. J. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (7763–7772) (2019).
5. Deng, J., Guo, J., Xue, N. & Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (4690–4699) (2019).
6. Deng, J., Guo, J., Ververas, E., Kotsia, I. & Zafeiriou, S. Retinaface: single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (5203–5212) (2020).
7. Wolf, L., Hassner, T. & Maoz, I. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (529–534) (2011).
8. Todt, J., Hanisch, S. & Strufe, T. Fantômas: evaluating reversibility of face anonymizations using a general deep learning attacker. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.10651 (2022).
9. Tournier, A. J. & De Montjoye, Y. A. Expanding the attack surface: robust profiling attacks threaten the privacy of sparse behavioral data. *Sci. Adv.* **8**, eabl6464 (2022).
10. European Union. General Data Protection Regulation 2016/679. https://www.privacy-regulation.eu/en/recital-26-GDPR.htm (2016). Accessed 21 April 2023.
11. Creemers, R., & Webster, G. Translation: personal information protection law of the People's Republic of China. *DigiChina Project* (2021).
12. Anderson, R. (2001). *Security Engineering: a Guide to Building Dependable Distributed Systems*. 240 (John Wiley & Sons, 2001).
13. Elliott, A. Better, broader, safer: using health data for research and analysis (the Goldacre review). *J. Radiol. Prot.* https://doi.org/10.1088/1361-6498/ac89f8 (2022).

Reporting summary

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

Data availability

The YouTube Faces Database is a publicly available dataset; for access, refer to ref. 7. The code and/or instructions for the replication of Yang et al.'s results as well as ours are available at https://github.com/computationalprivacy/unmask/.

Acknowledgements

We thank the authors of Yang et al. for helping us replicate their results.

Author contributions

M.M., S.J. and Y.-A.d.M. designed the experiments and wrote the paper. M.M. performed the experiments. M.M. and S.J. analyzed the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-023-02439-9.

Correspondence and requests for materials should be addressed to Yves-Alexandre de Montjoye.

Peer review information *Nature Medicine* thanks Juan Matias Di Martino and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Dr Yves-Alexandre de Montjoye

Last updated by author(s): Apr 25, 2023

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	We used the open source code FacePose_pytorch (https://github.com/WIKI2020/FacePose_pytorch) for the data collection process, specifically used to filter out images with an undesired pose
Data analysis	For the data analysis process, we used the following open source libraries : - RingNet (https://github.com/soubhiksanyal/RingNet): used to create the facial 2D masks given an image, based on the FLAME model (https://flame.is.tue.mpg.de/) - InsightFace (https://insightface.ai/): used their implementation of RetinaFace and ArcFace for face detection and recognition.

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

We use the Faces Database collected by this paper (https://ieeexplore.ieee.org/document/5995566) and made publicly available on their website (https://www.cs.tau.ac.il/~wolf/ytfaces/).

Human research participants

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

Reporting on sex and gender	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

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	We analyzed images of 555 unique individuals. This sample size has been determined by the maximum availability of unique facial images (after data exclusion) and matches the order of magnitude of the reference sample size of 405 individuals as used in Yang et al.
Data exclusions	As we want comparable quality of facial images as captured in the lab setting of Yang et al, we decided to only consider images where exactly one face was detected and where the estimated pitch and yaw were below 20 degrees (determined by FacePosePytorch).
Replication	All code used in the analysis is deterministic, except for the random sampling of facial images used for facial recognition. For the sampling we use a random seed, which we will provide, making the analysis fully reproducible.
Randomization	We did not allocate to any experimental groups - not relevant for our manuscript.
Blinding	Not relevant for our manuscript.

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