scientific data

Check for updates

OPEN

DATA DESCRIPTOR

PhotIQA: A photoacoustic image data set with image quality ratings

Anna Breger[1,2] ✉, Janek Gröhl[3,4,5], Clemens Karner[2], Thomas R. Else[3,4], Ian Selby[6,7], Tom Rix[8,9], Lara-Sophie Witt[5], Merle Duchêne[5], Jonathan Weir-McCall[6,10] & Carola-Bibiane Schönlieb[1]

Image quality assessment (IQA) is crucial in the evaluation stage of novel algorithms operating on images, including traditional and machine learning based methods. Due to the lack of available quality-rated medical images, most commonly used full-reference IQA measures have been developed and tested for natural images. Reported pitfalls and inconsistencies arising when applying such measures for medical images are not surprising, as they rely on different properties than natural images. In photoacoustic imaging (PAI), especially, standard benchmarking approaches for assessing the quality of image reconstructions are lacking. PAI is a multi-physics imaging modality, in which two inverse problems have to be solved, which makes the application of IQA measures uniquely challenging due to both, acoustic and optical, artifacts. To support the development and testing of IQA measures we assembled PhotIQA, a data set consisting of 1134 photoacoustic images. The images were rated by five experts across five quality properties in a full-reference setting, where the detailed rating enables usage beyond PAI. The data set with the images and corresponding ratings is publicly available on Zenodo.

Background

Advances in medical imaging technologies have been groundbreaking in the last decades, including the rapid development of deep learning methods that operate on a huge amount of image data. To ensure the feasibility of a novel methodology, quantitative image quality assessment (IQA) plays an important role for quality assurance in addition to visual inspection by experts, which is often limited due to time constraints, and therefore IQA measures may even serve as the main assessment criterion. Quantitative IQA measures may be distinguished based on the required information for the assessment step[1,2]. In full-reference (FR) IQA, a reference image is used to evaluate the quality of a corresponding (often degraded) image in a comparative way, relying on a meaningful notion of distance between the two images. In comparison, no-reference (NR) IQA aims to judge the quality without a reference based on predefined properties.

Many commonly used IQA measures have been developed for natural images and tested for specific tasks on a small number of publicly available, manually rated data sets. It is unknown how well these measures expand to medical images since they hold very distinct properties, and, moreover, often a different target space (color versus grayscale). Recent research has led to first insights on the applicability of common FR IQA measures to medical imaging data, see e.g.[3,4], showing incompatibilities of popular IQA measures to medical tasks. The research field suffers from the lack of publicly available image data sets with expert annotations, as well as unavailable implementations of published IQA measures. This is reflected in studies that show limitations in the design, e.g., relying on non-expert ratings (see e.g.[5]), non-realistic distortions (such as Gaussian additive noise, see e.g.[6]) or a very limited choice of IQA measures (see e.g.[7]). Combined with the scarce time resources of medical experts, there are many obstacles for the design of reproducible IQA comparison studies.

A first step to overcome these constraints is the design and sharing of realistic medical image data sets with expert quality ratings. In the NR setting, recently, there have been first data sets published, see e.g. a low-dose computed tomography data set[8]. The data set PhotIQA[9] is, to the best of our knowledge, the first publicly

[1]University of Cambridge, DAMTP, Cambridge, United Kingdom. [2]Medical University of Vienna, CMPBE, Vienna, Austria. [3]University of Cambridge, Department of Physics, Cambridge, UK. [4]Cancer Research UK Cambridge Institute, Cambridge, UK. [5]ENI-G, a Joint Initiative of the University Medical Center Göttingen and the Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany. [6]University of Cambridge, Department of Radiology, Cambridge, UK. [7]Cambridge University Hospitals, Department of Radiology, Cambridge, UK. [8]German Cancer Research Center (DKFZ), IMSY, Heidelberg, Germany. [9]Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany. [10]King's College London, School of Biomedical Engineering & Imaging Sciences, London, UK. ✉e-mail: ab2864@cam.ac.uk

available, open access data set with medical images rated by experts regarding quality in the FR setting, and specifically, for PAI. Note that due to the nature of FR IQA, the quality ratings may also be employed to assess NR IQA measures when discarding the reference images.

Photoacoustic imaging and IQA. Photoacoustic imaging (PAI) is an emerging medical imaging modality with important clinical applications, such as inflammatory bowel disease, cardiovascular diseases, and breast cancer[10]. It is a multi-physics modality, combining interactions of both light and sound with tissue. To reconstruct an image of the parameter of interest in PAI - the optical absorption coefficient - two inverse problems have to be solved: the acoustic inverse problem to reconstruct an image of absorbed energy density, and the optical inverse problem of recovering optical absorption, scattering, and the Grüneisen parameter. Through limitations in the measurement hardware and non-uniqueness of absorption and scattering, both inverse problems are not trivial to solve, and PAI thus suffers a variety of artifacts that can significantly affect the interpretability and clinical utility of the images[11,12].

The inverse problems of PAI pertain to accurately visualizing molecular distributions and determining functional tissue information from PA time series measurements[13]. It is important to identify suitable measures that can objectively assess the quality of PA image reconstructions, as finding the optimal reconstruction algorithm for the acoustic inverse problem given a certain hardware configuration is crucial for all further steps.

The state of the art is the employment of common quality measures, such as SSIM[1] or PSNR[14], which have on the other hand been shown to not act accurately for many medical image cases[15]. Especially in PAI, with multiple possible device configurations and diverse clinical applications[16], targeted IQA measures must be developed that can accurately assess the image quality given the diagnostic parameter to be extracted from the images. The quality-rated data we present in this paper can be an important first step for initiating experiments that validate the suitability of employed IQA measures. Moreover, the provided detailed quality properties allow a comprehensive insight into task-dependent suitability. As these properties are transferable to other medical imaging tasks, the ratings can also be used for assessing the suitability of IQA measures beyond PA images.

Methods

In this section, we will provide an overview of **PhotIQA**, the employed image data, and the newly obtained quality ratings.

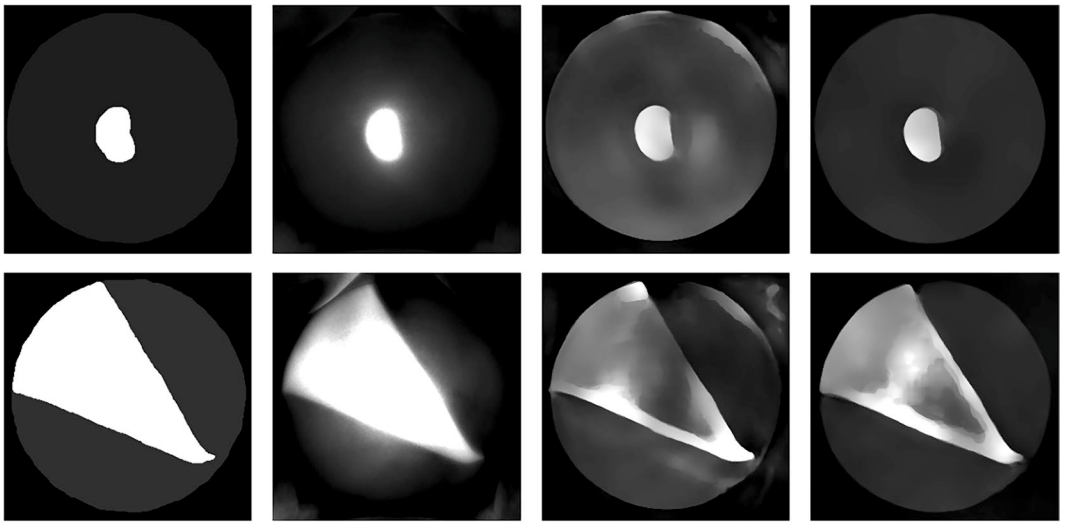
Photoacoustic Image Data. To obtain PhotIQA, we employed a previously published open access data set, cf.[17], available via[18], that consists of reconstructed PA images containing estimated distributions of the optical absorption coefficient from cross-sectional PA images of piecewise constant test objects (phantoms). The PA data were acquired with a preclinical commercial PA imaging system (MSOT InVision 256-TF, iThera Medical GmbH, Munich, Germany).

It contains 378 reference images that have been obtained using a double-integrating sphere[19] setup as a complementary measurement system, which yields point estimates for homogeneous material samples. Because of the piecewise-constant nature of the used phantoms, one can fabricate an additional batch of the material used for the test object, measure it, and relate the calculated properties to the test object. This process is unfeasible for complicated objects or in vivo images, but can serve in this setting to obtain reference images.

By applying 3 different reconstruction methods to the time series data yielding the spatial absorption coefficient images (see Fig. 1), we obtain 1134 reconstructed images, each corresponding to a reference image. The reconstruction method is a two-stage approach, approximating the acoustic inverse problem with a filtered back-projection and then applying one of three methods for the optical inverse problem. These methods were recently published[17] and can be summarized as follows: *Algorithm 1* is a fluence compensation algorithm, where the reconstructed image is corrected by a Monte Carlo simulation of the light fluence. The reference absorption and scattering values are obtained through double integrating sphere reference measurements, and this approach is only feasible here through the piecewise-constant nature of the phantoms. *Algorithm 2* is a deep learning algorithm trained in a supervised fashion on purely simulated data, and *Algorithm 3* is architecturally identical to the latter, but trained on experimentally acquired data. The deep learning architecture used for this case was a U-Net, modified for a regression task. The algorithms are described in detail in the publication by Gröhl *et al.*[17].

Image Quality Ratings. Five experts, with backgrounds in photoacoustic imaging, medical computer science, and medical physics, have been consulted to rate the reconstructed image data described above consulting a Likert scale, in particular, choosing 1 (very poor), 2 (poor), 3 (good) or 4 (very good) to describe the quality in comparison to the reference image. The rating was done by all experts on 2 days, taking approximately 9 hours per annotator. They were collected with the publicly available speedyIQA annotation app[20]. The software asks the user to set a task and rating categories, see Fig. 2. The experts were told not to change the contrast or luminance on their screen during the task. Obtained ratings have been saved in a CSV file, and to account for different rating behaviors between multiple graders, additionally, the z-score has been computed, cf.[21]. Eventually, the mean opinion scores (MOS) between the experts' ratings were computed. In the context of PhotIQA, the experts were asked to rate the images regarding

- Overall Quality,
- Edge Visibility (referred to as Edges),
- Object Homogeneity (referred to as Homogeneity),
- Object Intensity - Inclusion Area (referred to as Intensity_I), and
- Object Intensity - Background (referred to as Intensity_{BG}).



(a) Reference (b) *Algorithm 1* (c) *Algorithm 2* (d) *Algorithm 3*

Fig. 1 Two examples of images in PhotIQA, references (**a**) and the reconstructions from the described algorithms (**b**-**d**). *Algorithm 1* (**b**) corrects a reconstructed PA image by using the light fluence obtained from simulations. *Algorithm 2* and *Algorithm 3* (**c**,**d**) are deep-learning models trained to estimate the absorption coefficient.

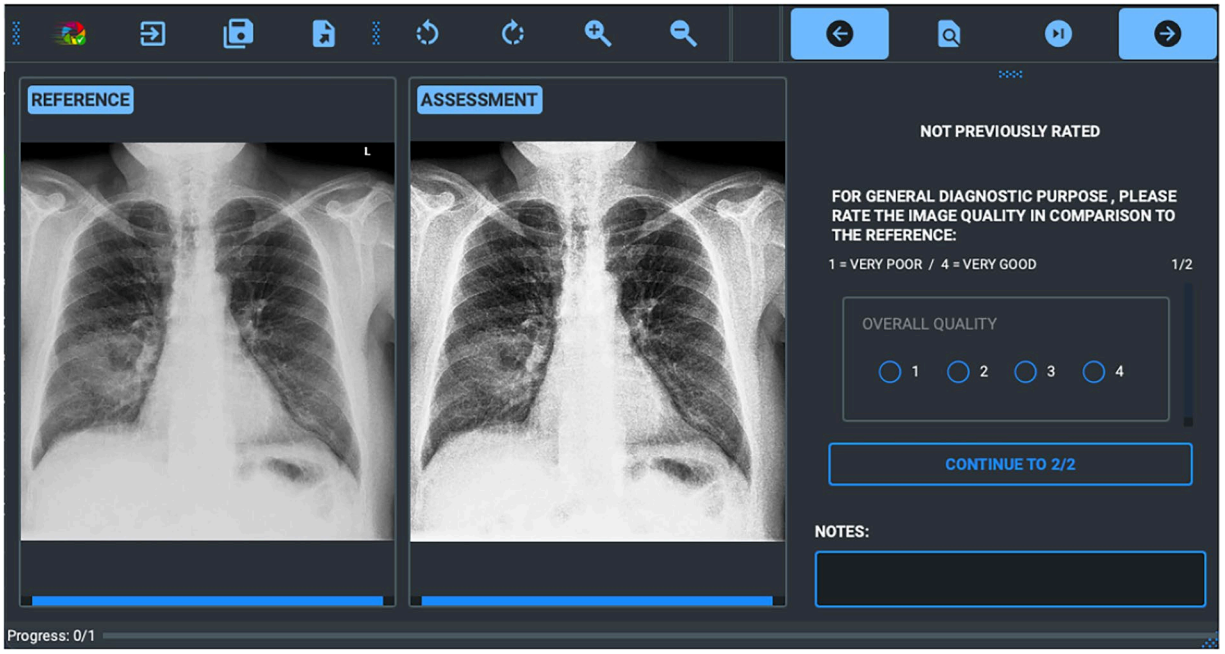


Fig. 2 The speedyIQA annotation app allows setting a task and rating categories for manual image quality ratings.

Here, the object intensity refers to any signals arising from inside the circular test object. Inclusion areas are those areas with increased contrast, and the background refers to the base material that makes up the majority of the test object. See Fig. 3 for an example of rated images according to the chosen properties.

Data Records

The PA images (378 references and 1134 corresponding reconstructions) and the image quality ratings are publicly available at (https://doi.org/10.5281/zenodo.13325196). The original PA image data is available via the University of Cambridge Apollo repository at (https://doi.org/10.17863/CAM.96644) and was published by Gröhl *et al.*[17] in 2023.

The PA images are stored as 8-bit *png* files of size 288 × 288 pixels in the *zip* file "*normalised_images.zip*". The image normalisation was done by rescaling all pixel values linearly such that the minimum is mapped to

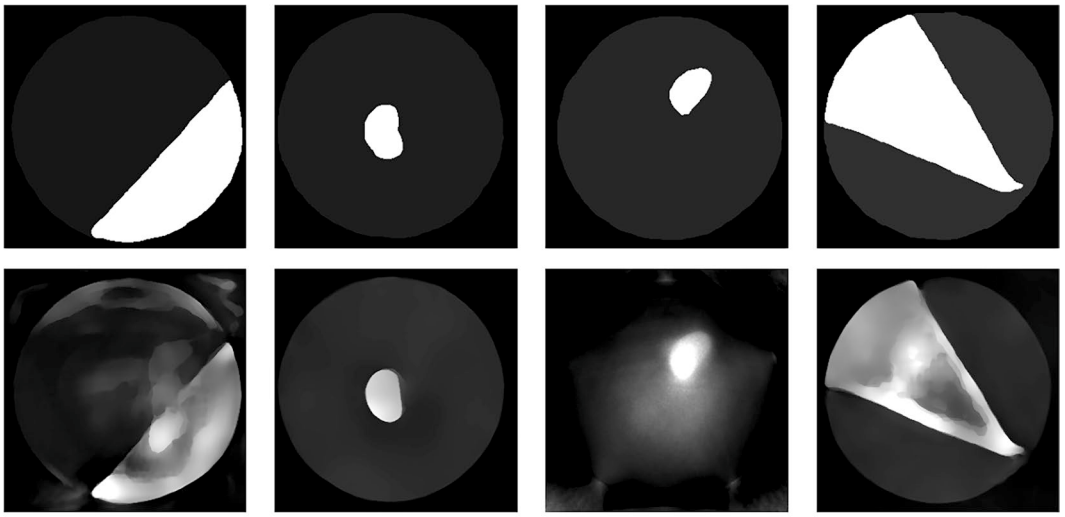


Fig. 3 Four examples of the minimum and maximum quality ratings over all annotators for the different detailed quality properties. (Top) reference image and (bottom) reconstructed, assessed image.

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Mean
Overall Quality	0.86	0.86	0.90	0.86	0.89	0.88
Edges	0.87	0.87	0.83	0.87	0.90	0.87
Homogeneity	0.90	0.90	0.88	0.83	0.87	0.88
Intensity$_I$	0.86	0.83	0.79	0.84	0.85	0.83
Intensity$_{BG}$	0.85	0.84	0.92	0.88	0.89	0.88
Mean	0.87	0.86	0.87	0.85	0.88	

Table 1. The table displays the Spearman Rank Correlation Coefficient (SRCC) for each task and rater, calculated between each rater's annotations and the z-scored MOS. The last column states the average SRCC value for each task. The last row states the average SRCC value for each rater.

0 and the maximum to 255. The minimum and maximum were defined based on the reference image, i.e., all values that were produced by an algorithm and exceeded the maximum of the reference were mapped to 255. This was done to streamline the use of all IQA measures while maintaining quantitative comparability. This *zip* file contains the folders "*./reference/*", all reference images named "*image*"+#IMAGE+"*.png*", and the folder "*./algorithms/*", three distorted images for each reference image named "*image*"+#IMAGE+"_"+#ALGORITHM+"*.png*" corresponding to the three employed algorithms described above.

The ratings of each quality property are stored in the file *annotations.zip*, which contains *csv* files named "*annotations_*"+*TASK*. These files contain the columns

- "filename": File name of the distorted image file
- "rater1",…, "rater5": Ratings from the 5 experts
- "mos": MOS of all 5 ratings
- "z-score rater1",…, "z-score rater5": Z-scores of all 5 ratings
- "z-score mos": MOS of the z-scores of all 5 ratings

The file "*IQA_evaluation.zip*" contains the script "*main.ipynb*", which allows to directly reproduce results of Table 2, as well as the text file "*software_versions.txt*", a list of the versions of all employed IQA metrics.

Technical Validation

Inter-rater consistency. First, we study the IQA ratings provided by the experts. In Table 1 and Fig. 4 we verify the agreement of ratings among them. The results in the table demonstrate a high correlation between all raters and tasks, and the latter shows the box plot of the absolute differences between the raters, for the raw data as well as the z-scores. We can see directly that the mean and median absolute difference of the raw ratings is at most 1, indicating a high alignment between the experts. The z-scores compensate for biases of individual raters

	Overall Quality	Edges	Homogeneity	Intensity$_I$	Intensity$_{BG}$
PSNR*	0.79/0.59	0.63/0.45	0.82/0.62	**0.74/0.56**	0.69/0.51
SSIM*	0.70/0.53	0.67/0.51	0.65/0.47	0.46/0.32	0.72/0.53
HaarPSI$_{med}$	**0.88/0.71**	**0.82/0.63**	0.83/0.63	0.60/0.42	**0.86/0.66**
HaarPSI	**0.87/0.68**	0.79/0.59	0.83/0.62	0.63/0.45	0.81/0.61
MS-SSIM*	**0.87/0.70**	0.74/0.56	**0.86/0.67**	0.59/0.41	**0.89/0.72**
MS-SSIM	**0.91/0.74**	**0.80/0.61**	**0.89/0.71**	**0.67/0.48**	**0.88/0.71**
IW-SSIM*	**0.87/0.69**	0.77/0.58	**0.85/0.65**	0.65/0.46	**0.86/0.67**
GMSD*	0.83/0.64	0.68/0.49	**0.85/0.67**	**0.68/0.49**	0.77/0.58
FSIM*	0.83/0.65	0.70/0.52	0.84/0.64	0.64/0.45	0.82/0.62
MDSI*	0.75/0.55	0.65/0.47	0.74/0.54	0.66/0.47	0.65/0.46
LPIPS$_{Alex}$	0.82/0.63	**0.81/0.63**	0.74/0.55	0.47/0.32	**0.86/0.67**
DISTS*	0.76/0.59	0.67/0.50	0.75/0.57	0.45/0.32	0.82/0.64
NIQE*	0.56/0.39	0.34/0.23	0.70/0.51	0.53/0.38	0.54/0.38
PaQ-2-PiQ	0.30/0.20	0.51/0.35	0.12/0.08	0.05/0.04	0.33/0.22
TeamEpoch	0.25/0.17	0.39/0.27	0.09/0.06	0.04/0.02	0.33/0.22
UIQA	0.52/0.35	0.58/0.40	0.43/0.28	0.25/0.17	0.48/0.33

Table 2. Baseline results: SRCC/KRCC between the MOS of the manual expert ratings' z-scores and common full-reference (top) and no-reference (bottom) IQA measures. *denotes that implementations were provided in MATLAB rather than Python. The 3 highest results are bold.

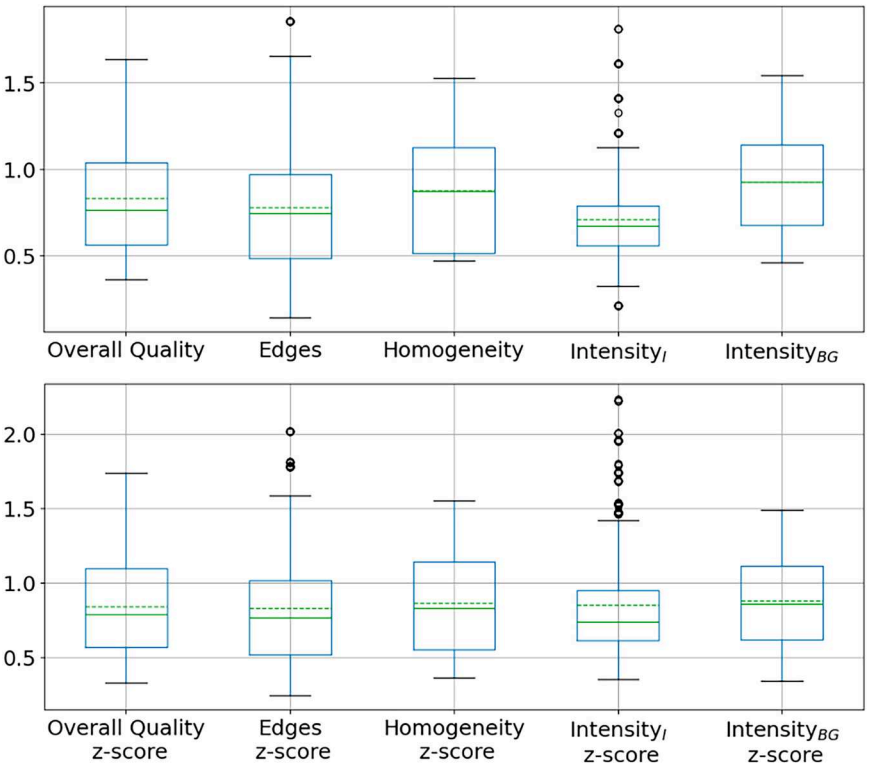


Fig. 4 Box plot of the mean absolute differences (top) and the absolute differences of z-scores (bottom) of all 5 raters to the (z-scored) MOS with the median (green line) and mean (striped green line).

towards low or high scores that may limit the direct interpretability. For example, the meaning of the rating (4) might differ between the experts depending on a tendency towards low or high ratings, and the z-scores allow insights into the corrected difference. In the box plot corresponding to the z-scores, we can observe that the agreement acts similarly across the properties. The categories "Intensity$_I$" and "Edges" contain the most outliers. Visual examples of extreme outliers in those categories are provided in Fig. 6.

Moreover, in Fig. 5, we show the amount of each rating given per task. We observe that the rating (1) was given much less than the ratings (2-4), which is caused by the reconstructed images stemming from three reconstruction algorithms that all give somewhat reasonable results. The property "Intensity$_I$" peaks at the highest rating (4). The other properties show behavior closer to a normal distribution peaking at the rating (3). Rater 3

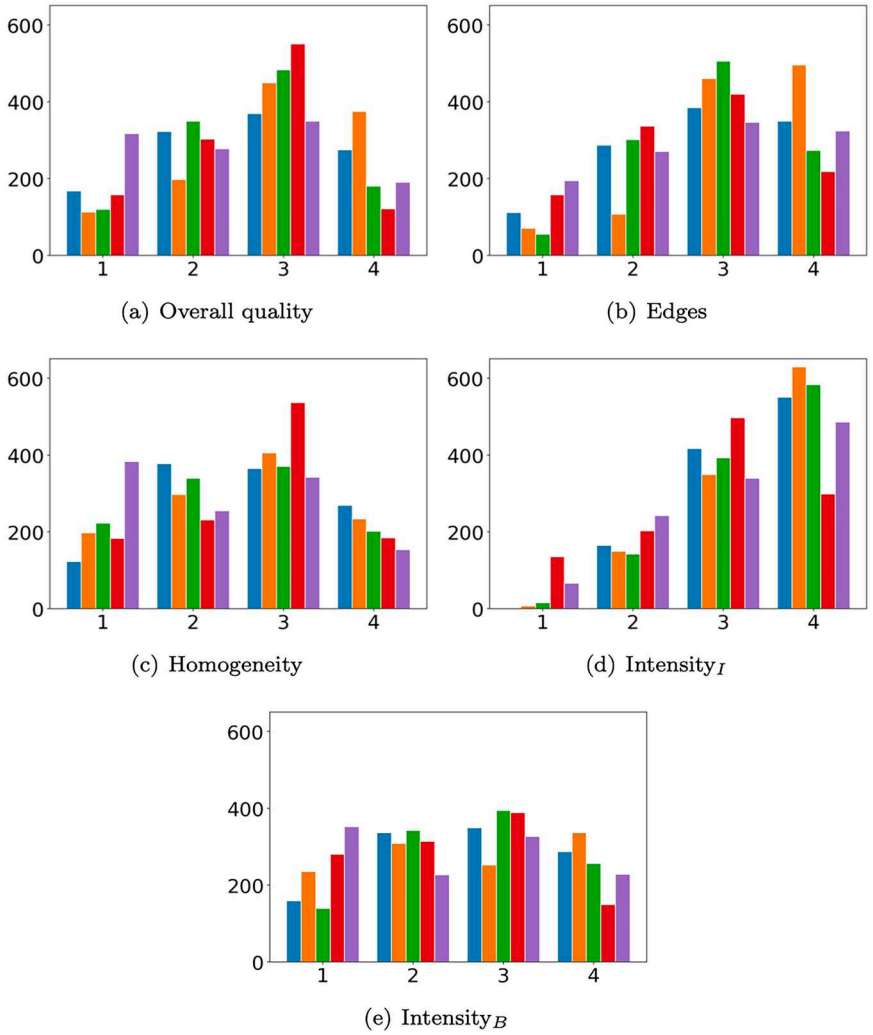


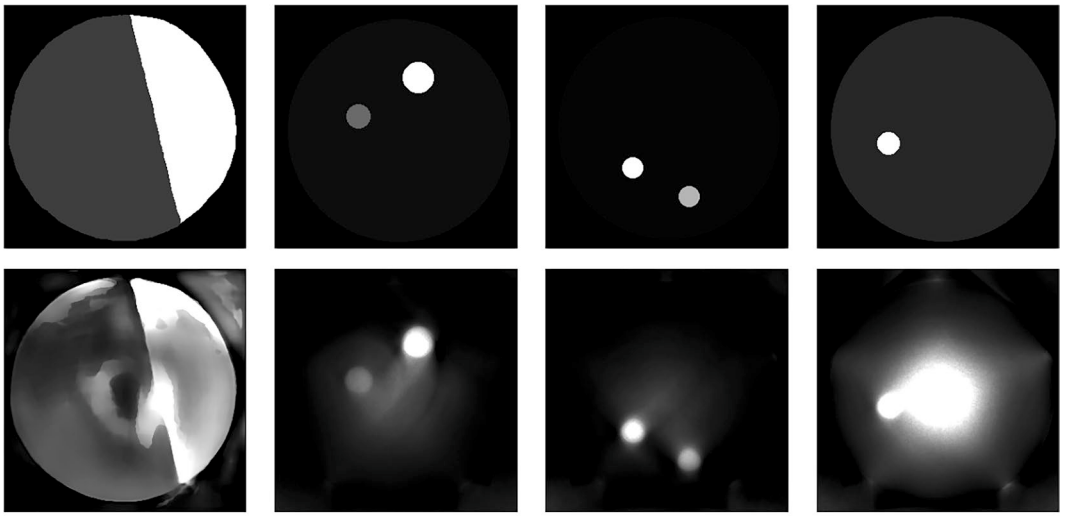
Fig. 5 The distribution of the ratings for each quality property (**a**–**e**) plotted for the 5 annotators.

showed a different rating behavior regarding (4), which they used much less than the other raters, and rater 2 leaned towards higher ratings. The z-score accounts for those personal tendencies.

IQA evaluation. To provide a baseline assessment of the validity of common IQA measures, we compare their values to the manual ratings. The chosen IQA measures were computed whenever possible with implementations provided by their authors using default parameters, either in MATLAB or Python, and in the case for PSNR, SSIM and MS-SSIM, we utilise commonly used Python and Matlab implementations in their default settings. The exact versions and software packages are stated as part of the data set on Zenodo, see Section *Data Records* above for more details.

For the evaluation of the IQA measures' performance, we employ the Spearman Rank Correlation Coefficient (SRCC) and the Kendall Rank Correlation Coefficient (KRCC) on the whole image dataset, which assess the rank between the IQA measures and the manual ratings. The absolute SRCC and KRCC between the MOS of the z-scores and the IQA measures are stated in Table 2. We can observe that MS-SSIM[22], with the implementation from the Python torchmetrics library, is among the top 3 for all properties. It is important to notice that there is a relevant difference to the implementation provided in MATLAB. In line with previous experiments, HaarPSI[23], especially in the HaarPSI$_{\text{med}}$[24] setting, as well as IW-SSIM[25], yield also good results in all categories. LPIPS[26] and GMSD[27] show promising behavior for some properties. The commonly used measures PSNR and SSIM struggle for almost all quality properties. PSNR is the best measure regarding "Intensity$_I$", which is not surprising, as it directly computes the pixel-wise difference and is sensitive to shifts in high intensity values.

Other FR IQA measures include FSIM[28], MDSI[29] and DISTS[30]. The tested NR measures NIQE[31], PaQ-2-PiQ[32], TeamEpoch[8] and UIQA[33] yield generally lower results, which is not surprising, as this a more distinct and different task than the direct comparison to a reference image. It has also to be noted that the images were annotated in a FR and not NR setting, positively biasing the evaluation for FR measures.



Overall Quality:	1|2	Overall Quality:	1|3	Overall Quality:	2|3	Overall Quality:	1|1
Edges:	1|4	**Edges:**	1|4	Edges:	1|3	Edges:	1|1
Homogeneity	1|2	Homogeneity	1|3	Homogeneity:	2|3	Homogeneity:	1|2
Intensity_I	2|3	Intensity_I	3|4	**Intensity_I:**	1|4	**Intensity_I:**	1|4
Intensity_{BG}	1|2	Intensity_{BG}	2|2	Intensity_{BG}:	2|3	Intensity_{BG}:	1|1

Fig. 6 Four examples of the minimum and maximum quality ratings from all annotators corresponding to box plot outliers in Fig. 4 (properties in bold). Reference image (top) and reconstructed, assessed image (bottom).

Usage Notes

In order to pair the images with the scores of a specific property, e.g. overall quality, we employ the *csv* file "*annotations_overall_quality.csv*". The filename leads to the corresponding *png* images in the folders "*./algorithms/*" and "*./references/*" contained in the *zip* file "*normalised_images.zip*".

Limitations. The ratings were obtained by 5 experts and, as expected for manual annotations, show some variability in rating behavior, see Fig. 6. Nevertheless, we also show in Table 1 and Fig. 4, that, overall, the ratings correlate highly between all raters.

The two-stage PA reconstruction was done using filtered back-projection for the acoustic inverse problem. While we expect a deep learning based approach for the optical inversion to learn to account for systematic reconstruction artifacts, the concrete algorithm for the acoustic inversion is expected to have an impact[34]. Moreover, in this work, we evaluate the PAI reconstruction performance purely on the basis of comparing reconstructed absorption coefficients to reference absorption coefficients. In a clinical setting, this might not always be the most important endpoint. Furthermore, the narrow endpoint limits the spectrum of PAI artifacts represented by this dataset.

A major benefit of the dataset is the access to known underlying tissue properties obtained with established reference measurement methods, which allows the benchmarking of FR IQA measures. However, the clinical meaningfulness is limited; on a clinical dataset, one would not have access to such ground-truth knowledge. This hinders the use of FR image quality measures. If the true image reconstruction is unknown, this leaves only NR (or reduced-reference) measures. Such measures could then be benchmarked against medical expert annotators as well as clinical outcome measures.

We normalise the images between 0 and 255 given the bounds of the reference image. While this streamlines the application of the IQA measures, it simultaneously introduces a bias, where systematic over/underestimations of the algorithms may be masked. Because of this reason, we did not introduce classical quantitative distance measures, such as the root mean squared error or absolute/relative error measures, and suggest computing such measures on the unnormalised images available in the original data publication[17,18].

Data availability

The normalised images and manual quality ratings are publicly available under a Creative Commons license (CC-BY 4.0) at https://doi.org/10.5281/zenodo.13325196[9]. The original phantom data is publicly available under a Creative Commons license (CC-BY 4.0) at https://doi.org/10.17863/CAM.96644[18].

Code availability

The code for the PAI reconstruction algorithms is available open source at (https://github.com/BohndiekLab/end_to_end_phantom_QPAT) under an MIT license. The Python code of the IQA experiments, including the implementations of the Python based IQA measures, is available on GitHub https://github.com/ideal-iqa/iqa-eval under an MIT license. A Python file to reproduce results in Table 2 is available under a Creative Commons license

(CC-BY 4.0) at https://doi.org/10.5281/zenodo.13325196[9]. The speedyIQA annotation app is available at https://github.com/selbs/speedy_iqa under an MIT license.

Received: 31 July 2025; Accepted: 2 June 2026;
Published online: 17 June 2026

References

1. Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612, https://doi.org/10.1109/TIP.2003.819861 (2004).
2. Athar, S. & Wang, Z. A comprehensive performance evaluation of image quality assessment algorithms. *IEEE Access* **7**, 140030–140070, https://doi.org/10.1109/ACCESS.2019.2943319 (2019).
3. Kastryulin, S., Zakirov, J., Pezzotti, N. & Dylov, D. V. Image quality assessment for magnetic resonance imaging. *IEEE Access* **11**, 14154–14168, https://doi.org/10.1109/ACCESS.2023.3243466 (2023).
4. Breger, A. *et al*. A study on the adequacy of common iqa measures for medical images. In *Proceedings of 2024 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD), Springer Lecture Notes in Electrical Engineering* (2024).
5. Chow, L. S., Rajagopal, H. & Paramesran, R. Correlation between subjective and objective assessment of magnetic resonance (mr) images. *Magn Reson Imaging* **34**(6), 820–831, https://doi.org/10.1016/j.mri.2016.03.006 (2016).
6. Ohashi, K. *et al*. Applicability evaluation of full-reference image quality assessment methods for computed tomography images. *Journal of Digital Imaging* **36**(6), 2623–2634, https://doi.org/10.1007/s10278-023-00875-0 (2023).
7. Renieblas, G. P., Nogués, A. T., González, A. M., Gómez-Leon, N. & Del Castillo, E. G. Structural similarity index family for image quality assessment in radiological images. *J Med Imaging (Bellingham)* **4**(3), 035501, https://doi.org/10.1117/1.JMI.4.3.035501 (2017).
8. Lee, W. *et al*. Low-dose computed tomography perceptual image quality assessment. *Medical Image Analysis* **99**, 103343, https://doi.org/10.1016/j.media.2024.103343 (2025).
9. Breger, A. *et al*. Photiqa: A photoacoustic image data set with image quality ratings. *Zenodo* https://doi.org/10.5281/zenodo.13325196 (2025).
10. Assi, H. *et al*. A review of a strategic roadmapping exercise to advance clinical translation of photoacoustic imaging: From current barriers to future adoption. *Photoacoustics* **32**, 100539, https://doi.org/10.1016/j.pacs.2023.100539 (2023).
11. Rietberg, M. T. *et al*. Artifacts in photoacoustic imaging: Origins and mitigations. *Photoacoustics* **45**, 100745, https://doi.org/10.1016/j.pacs.2025.100745 (2025).
12. Else, T. R. *et al*. The confounding effects of skin colour in photoacoustic imaging. *medRxiv* 2025–03 (2025).
13. Cox, B., Laufer, J. G., Arridge, S. R. & Beard, P. C. Quantitative spectroscopic photoacoustic imaging: a review. *J Biomed Opt* **17**(6), 061202, https://doi.org/10.1117/1.JBO.17.6.061202 (2012).
14. Gröhl, J., Schellenberg, M., Dreher, K. & Maier-Hein, L. Deep learning for biomedical photoacoustic imaging: A review. *Photoacoustics* **22**, 100241 (2021).
15. Breger, A. *et al*. A study of why we need to reassess full reference image quality assessment with medical images. *Journal of Imaging Informatics in Medicine*https://doi.org/10.1007/s10278-025-01462-1 (2025).
16. Park, J. *et al*. Clinical translation of photoacoustic imaging. *Nature Reviews Bioengineering* **3**(3), 193–212, https://doi.org/10.1038/s44222-024-00240-y (2025).
17. Gröhl, J. *et al*. Moving beyond simulation: data-driven quantitative photoacoustic imaging using tissue-mimicking phantoms. *IEEE Trans Med Imaging* **PP**, https://doi.org/10.1109/TMI.2023.3331198 (2023).
18. Gröhl, J. *et al*. Dataset for: Moving beyond simulation: data-driven quantitative photoacoustic imaging using tissue-mimicking phantoms. https://doi.org/10.17863/CAM.96644, https://www.repository.cam.ac.uk/handle/1810/359968 (2023).
19. Pickering, J. W. *et al*. Double-integrating-sphere system for measuring the optical properties of tissue. *Appl. Opt.* **32**(4), 399–410, https://doi.org/10.1364/AO.32.000399 (1993).
20. Selby, I. Github repository speedyiqa https://github.com/selbs/speedy_iqa (March 2024).
21. Sheikh, H., Sabir, M. & Bovik, A. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing* **15**(11), 3440–3451, https://doi.org/10.1109/TIP.2006.881959 (2006).
22. Zhou Wang, E. P. S. & Bovik, A. C. Multi-scale structural similarity for image quality assessment. In *Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA* (2003).
23. Reisenhofer, R., Bosse, S., Kutyniok, G. & Wiegand, T. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Process. Image Commun.* **61**, 33–43, https://doi.org/10.1016/j.image.2017.11.001 (2018).
24. Karner, C. *et al*. Parameter choices in haarpsi for iqa with medical images. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–5, https://doi.org/10.1109/ISBI60581.2025.10981227 (2025).
25. Wang, Z. & Li, Q. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing* **20**(5), 1185–1198, https://doi.org/10.1109/TIP.2010.2092435 (2011).
26. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595, https://doi.org/10.1109/CVPR.2018.00068 (2018).
27. Xue, W., Zhang, L., Mou, X. & Bovik, A. C. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing* **23**(2), 684–695, https://doi.org/10.1109/TIP.2013.2293423 (2014).
28. Zhang, L., Zhang, L., Mou, X. & Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing* **20**(8), 2378–2386, https://doi.org/10.1109/TIP.2011.2109730 (2011).
29. Ziaei Nafchi, H., Shahkolaei, A., Hedjam, R. & Cheriet, M. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access* **4**, 5579–5590, https://doi.org/10.1109/ACCESS.2016.2604042 (2016).
30. Ding, K., Ma, K., Wang, S. & Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(5), 2567–2581, https://doi.org/10.1109/TPAMI.2020.3045810 (2022).
31. Mittal, A., Soundararajan, R. & Bovik, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters* **20**(3), 209–212, https://doi.org/10.1109/LSP.2012.2227726 (2013).
32. Ying, Z. *et al*. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. 3572–3582, https://doi.org/10.1109/CVPR42600.2020.00363 (2020).
33. Sun, W. *et al*. Assessing uhd image quality from aesthetics, distortions, and saliency. In Del Bue, A., Canton, C., Pont-Tuset, J. & Tommasi, T. (eds.) *Computer Vision – ECCV 2024 Workshops*, 109–126 (Springer Nature Switzerland, Cham, 2025).
34. Gröhl, J. *et al*. Digital twins enable full-reference quality assessment of photoacoustic image reconstructions. *The Journal of the Acoustical Society of America* **158**(1), 590–601, https://doi.org/10.1121/10.0037188 (2025).

Acknowledgements

The authors wish to acknowledge support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking - DRAGON (101005122) (A.B., I.S., C.B.S.); the Austrian Science Fund (FWF) through project T1307 (A.B., C.K.);

the German Research Foundation through the grant GR 5824/1 (J.G.) and project number 462569370 (T.R.); EPSRC UK EP/X037770/1 (T.R.E.); Cancer Research UK through C9545/A29580 (T.R.E); the ERC under the European Union's Horizon research and innovation programme through project NEURAL SPICING (grant 101002198) (T.R.),the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) (I.S.) and (NIHR203312) (C.B.S); C.B.S also acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC (EP/V029428/1, EP/V026259/1, EP/S026045/1, EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the EPSRC funded ProbAI hub EP/Y028783/1), and the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement REMODEL. Please note that the content of this publication reflects the authors' views and that neither NIHR, the Department of Health and Social Care, IMI, the European Union, EFPIA, nor the DRAGON consortium are responsible for any use that may be made of the information contained therein.

Author contributions

A.B., J.G., T.R.E. and I.S. designed the study and performed the data collection; A.B. and J.G. wrote the initial version of the paper; J.G., T.R.E., T.R., L.-S.W. and M.D. performed data annotation; C.K. and A.B. worked on the technical validation and figures. All authors discussed the results and contributed to the editing of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.B.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026