IOP Publishing Phys. Med. Biol. 68 (2023) 074001 https://doi.org/10.1088/1361-6560/acc0ce

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
17 August 2022

REVISED
20 January 2023

ACCEPTED FOR PUBLICATION
2 March 2023

PUBLISHED
21 March 2023

PAPER

Observer-study-based approaches to quantitatively evaluate the realism of synthetic medical images

Ziping Liu[1], Scott Wolfe[2], Zitong Yu[1], Richard Laforest[2,3], Joyce C Mhlanga[2], Tyler J Fraum[2,3], Malak Itani[2], Farrokh Dehdashti[2,3], Barry A Siegel[2,3] and Abhinav K Jha[1,2,3,*]

[1] Department of Biomedical Engineering, Washington University, St. Louis, MO 63130, United States of America
[2] Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110, United States of America
[3] Alvin J. Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO 63110, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: a.jha@wustl.edu

Keywords: image synthesis, image quality assessment, medical imaging, observer study

Abstract

Objective. Synthetic images generated by simulation studies have a well-recognized role in developing and evaluating imaging systems and methods. However, for clinically relevant development and evaluation, the synthetic images must be clinically realistic and, ideally, have the same distribution as that of clinical images. Thus, mechanisms that can quantitatively evaluate this clinical realism and, ideally, the similarity in distributions of the real and synthetic images, are much needed. *Approach.* We investigated two observer-study-based approaches to quantitatively evaluate the clinical realism of synthetic images. In the first approach, we presented a theoretical formalism for the use of an ideal-observer study to quantitatively evaluate the similarity in distributions between the real and synthetic images. This theoretical formalism provides a direct relationship between the area under the receiver operating characteristic curve, AUC, for an ideal observer and the distributions of real and synthetic images. The second approach is based on the use of expert-human-observer studies to quantitatively evaluate the realism of synthetic images. In this approach, we developed a web-based software to conduct two-alternative forced-choice (2-AFC) experiments with expert human observers. The usability of this software was evaluated by conducting a system usability scale (SUS) survey with seven expert human readers and five observer-study designers. Further, we demonstrated the application of this software to evaluate a stochastic and physics-based image-synthesis technique for oncologic positron emission tomography (PET). In this evaluation, the 2-AFC study with our software was performed by six expert human readers, who were highly experienced in reading PET scans, with years of expertise ranging from 7 to 40 years (median: 12 years, average: 20.4 years). *Main results.* In the ideal-observer-study-based approach, we theoretically demonstrated that the AUC for an ideal observer can be expressed, to an excellent approximation, by the Bhattacharyya distance between the distributions of the real and synthetic images. This relationship shows that a decrease in the ideal-observer AUC indicates a decrease in the distance between the two image distributions. Moreover, a lower bound of ideal-observer AUC = 0.5 implies that the distributions of synthetic and real images exactly match. For the expert-human-observer-study-based approach, our software for performing the 2-AFC experiments is available at https://apps.mir.wustl.edu/twoafc. Results from the SUS survey demonstrate that the web application is very user friendly and accessible. As a secondary finding, evaluation of a stochastic and physics-based PET image-synthesis technique using our software showed that expert human readers had limited ability to distinguish the real images from the synthetic images. *Significance.* This work addresses the important need for mechanisms to quantitatively evaluate the clinical realism of synthetic images. The mathematical treatment in this paper shows that quantifying the similarity in the distribution of real and synthetic images is theoretically possible by using an ideal-observer-study-based approach. Our developed software

provides a platform for designing and performing 2-AFC experiments with human observers in a highly accessible, efficient, and secure manner. Additionally, our results on the evaluation of the stochastic and physics-based image-synthesis technique motivate the application of this technique to develop and evaluate a wide array of PET imaging methods.

1. Introduction

In medical imaging, the use of simulation studies to develop and objectively evaluate new and improved imaging methods has been well recognized (Frangi *et al* 2018, Abadi *et al* 2020, Jha *et al* 2021, 2022, Yousefirizi *et al* 2021). Simulation studies offer the advantage of evaluating the performance of a method against known ground truth, provide the ability to accurately model patient anatomy and physiology as well as imaging system characteristics, incorporate population variability, and generate multiple scan realizations of the same patient to evaluate reproducibility. Even more importantly, this is all done in silico, which is inexpensive and enables optimizing the method before conducting clinical studies. Given these advantages, simulation studies have been used to evaluate a wide range of imaging methods for system instrumentation (Surti *et al* 2006), image reconstruction (Song *et al* 2011), image enhancement (Yu *et al* 2020), and image segmentation (Liu *et al* 2022). Further, the advantages of simulation studies have led to the emergence of virtual clinical trial-based frameworks to evaluate imaging methods (Maidment 2014, Badano *et al* 2018, Abadi *et al* 2020, Badano 2021, Li *et al* 2022). Simulation studies have also shown promise in developing artificial intelligence (AI)-based algorithms for medical imaging. More specifically, a key challenge in developing AI-based algorithms is the requirement of large amounts of training data with known ground truth. This data can be difficult, expensive, and time-consuming to obtain, thus creating a barrier to developing learning-based algorithms. Studies have shown that synthetic images generated from simulations can help alleviate this requirement by providing such training data for purposes such as pre-training the network (Chartsias *et al* 2017a, Creswell *et al* 2018, Gong *et al* 2018, Guan and Loew 2019, Leung *et al* 2020).

For the simulation-based development and evaluation studies to yield clinically relevant inferences, it is important that images generated by the synthesis techniques are clinically realistic (Song *et al* 2011, Jha *et al* 2016, 2021). Ensuring this clinical realism requires that patient anatomy and physiology, population variability, and imaging-system physics are all modeled accurately. There has been much work on evaluating the accuracy in modeling the imaging physics (Gonias *et al* 2007, Poon *et al* 2015, Hernandez-Giron *et al* 2019). However, fewer studies have focused on developing approaches to ensure that the population variability is modeled accurately (Badano *et al* 2018, Zhou *et al* 2019a, Houbrechts *et al* 2021). Note that to ensure clinical realism, it is not sufficient to just assess whether the real and synthetic images match for one patient realization. Instead, for clinically relevant studies, the ideal goal is that the distributions of real and synthetic images should match. This provides confidence that the findings of objective evaluation studies with synthetic images, including virtual clinical trials, are clinically relevant. Further, the clinical realism of synthetic images has been observed to be necessary when using these images for pre-training AI-based algorithms (Leung *et al* 2020). Thus, there is an important need for mechanisms that can quantitatively evaluate the clinical realism of synthetic images and, ideally, the similarity in distributions of real and synthetic images. To address this need, we present two observer-study-based approaches in this manuscript, one based on the ideal observer and the other based on the human observer.

To quantify the distance between distributions of real and synthetic images, metrics such as the Fréchet inception distance (FID) (Heusel *et al* 2017) have been proposed. The FID measures the difference between the statistics extracted from real and synthetic images using a pre-trained Inception network. However, this network is typically pre-trained on ImageNet, which comprises only natural images. Thus, it is unclear whether the network can effectively generalize to evaluate the realism of synthetic medical images. Another set of metrics attempt to evaluate the difference between distributions of real and synthetic images based on the performance of an image classifier (Shmelkov *et al* 2018). These approaches, while promising, rely on the choice of the classifier. More importantly, it is theoretically unclear whether this performance relates to the similarity in distributions between the real and synthetic images.

More recently, observer-study-based approaches have been considered to evaluate the clinical realism of synthetic images (Burgess 2011, Chen *et al* 2016, Elangovan *et al* 2017, Ma *et al* 2017, Sturgeon *et al* 2017). In these approaches, a two-alternative forced-choice (2-AFC) experiment is typically performed. In this 2-AFC experiment, an observer is presented pairs of real and synthetic images. For each image pair, the observer is asked to identify the real image. It is well accepted that the probability of correctly identifying the real image is equivalent to the area under the receiver operating characteristics curve, AUC, for that observer (Barrett and Myers 2013). Thus, if an observer correctly identifies the real images for only 50% of the cases, this yields an AUC

of 0.5. Consequently, this implies that the observer is unable to differentiate the real images from the synthetic images. However, this does not necessarily indicate that the distribution of synthetic images matches that of real images. To illustrate this point, we consider a numerical observer. This observer, in the 2-AFC experiment, calculates a test statistic for each image and identifies the image that yields a higher value of test statistic as real. However, the test statistic is just a single statistic derived from the entire image. Thus, while an AUC of 0.5 may indicate that the distributions of the test statistic of the real and synthetic images match, this does not necessarily indicate that distributions of the real and synthetic images also match. Further, when the AUC value is greater than 0.5, it is unclear how the AUC value relates to the distance between the distributions of real and synthetic images. A mathematical analysis for answering these questions is much needed.

The first goal of this work is to theoretically demonstrate that an ideal observer provides a mechanism to quantify the similarity in distributions between the real and synthetic images. This ideal observer, also referred to as the likelihood-ratio test, uses all the statistical information available in the data to maximize task performance. Further, this observer is numerical and, thus, paves the way for a mathematical analysis. In this context, in 1998, Barrett *et al* (1998) published a seminal paper with the goal of bridging the gap between the use of signal-to-noise ratio and the use of the AUC as a figure of merit for signal-detection tasks. In that paper, one of the important findings was deriving the AUC for an ideal observer explicitly in terms of the distributions of signal-present and signal-absent images. By following a similar mathematical treatment as in Barrett *et al*, but in the context of evaluating the clinical realism of synthetic images, we show that an ideal-observer-study-based approach can be used to quantitatively assess the similarity in distributions of the real and synthetic images (section 2). Specifically, we show that the ideal-observer AUC is related, to an excellent approximation, to the Bhattacharyya distance (Bhattacharyya 1943) between the distributions of the real and synthetic images.

The second goal of this work is to develop an openly-available web-based platform to evaluate the clinical realism of synthetic images using human-observer studies. In this context, a vast majority of observer-study-based approaches to evaluate the clinical realism of synthetic images have relied on the use of human observers (Burgess 2011, Chen *et al* 2016, Elangovan *et al* 2017, Ma *et al* 2017, Sturgeon *et al* 2017). Among the different human observers, physicians have multiple years of experience reading medical images and are very familiar with the intricate details of these images. Thus, these physicians, whom we refer to as expert human observers, are best placed to identify even minute differences between the real and synthetic images. To conduct observer studies with expert human readers, various software have been developed. However, these software often require manual installation on local workstations with compatible operating systems (Håkansson *et al* 2010, Zhang *et al* 2016, Genske and Jahnke 2022). The variety in existing operating systems and the fact that users must obtain administrative privileges to install software on workstations owned by institution limit the accessibility of those software. Consequently, these factors make it challenging and cumbersome to conduct human-observer studies. Thus, an accessible and easy-to-use tool that can facilitate the conducting of expert-human-observer studies for evaluating the realism of synthetic images is much needed. Our developed web-based platform (section 3) is in the direction of addressing this need.

2. Ideal-observer-study-based approach to quantitatively evaluate the similarity in the distributions of real and synthetic images

2.1. Problem formulation

Consider a set of clinical images that are acquired from a population of patients scanned by a medical-imaging system. Denote the image of each patient by an M-dimensional vector, $\hat{\mathbf{f}}^r$, which, we assume, lies within the Hilbert space of Euclidean vectors, denoted by $\mathbb{E}^M$. Additionally, consider an image-synthesis method that generates images of a simulated population of patients in silico. Each synthetic medical image, denoted by an M-dimensional vector, $\hat{\mathbf{f}}^s$, is also assumed to lie within $\mathbb{E}^M$.

To evaluate the clinical realism of those synthetic images, we consider a 2-AFC experiment being performed by a numerical observer. In this experiment, an observer is presented with pairs of real and synthetic images, $\hat{\mathbf{f}}^r$ and $\hat{\mathbf{f}}^s$. The classes of synthetic and real images are referred to as the hypotheses H_1 and H_2, respectively. Denote the probability of observing an image $\hat{\mathbf{f}}$ under the hypothesis H_j by $\mathrm{pr}(\hat{\mathbf{f}}|H_j)$. Then, $\hat{\mathbf{f}}^s$ is sampled from $\mathrm{pr}(\hat{\mathbf{f}}|H_1)$ and $\hat{\mathbf{f}}^r$ is sampled from $\mathrm{pr}(\hat{\mathbf{f}}|H_2)$. The observer is then required to identify the real image. To make this decision, the observer calculates two test statistics, $\theta(\hat{\mathbf{f}}^s)$ and $\theta(\hat{\mathbf{f}}^r)$, and assigns the image that yields the higher value of the test statistic to H_2. The decision is correct if $\theta(\hat{\mathbf{f}}^r) > \theta(\hat{\mathbf{f}}^s)$. For convenience of notation, let $q_j(\hat{\mathbf{f}}) \equiv \mathrm{pr}(\hat{\mathbf{f}}|H_j)$. The probability of a correct decision can be calculated as

$$\Pr[\theta(\hat{\mathbf{f}}^r) > \theta(\hat{\mathbf{f}}^s)] = \int_\infty \mathrm{d}^M\hat{\mathbf{f}}^s \int_\infty \mathrm{d}^M\hat{\mathbf{f}}^r \, q_1(\hat{\mathbf{f}}^s) q_2(\hat{\mathbf{f}}^r) \, \mathrm{step}(\theta(\hat{\mathbf{f}}^r) - \theta(\hat{\mathbf{f}}^s)), \tag{1}$$

where step(·) denotes the Heaviside unit step function. As shown in Barrett and Myers (2013) in the context of signal-detection tasks and rephrased in this scenario of using the 2-AFC experiment to evaluate the clinical realism of synthetic images (appendix A), the right-hand side of the above expression is equivalent to the expression for the AUC for an observer in terms of integrals over $\hat{\mathbf{f}}^r$ and $\hat{\mathbf{f}}^s$. Thus, from equation (1), the accuracy of an observer in identifying the real images in a 2-AFC experiment is equivalent to the AUC for that observer.

We note that the expression for the AUC using equation (1) depends on the test statistics and, thus, does not specify a direct relationship between the AUC value and the distance between the distributions of the real and synthetic images. To gain insights into this relationship, we consider the use of an ideal observer, which uses all the statistical information available in the data to evaluate the realism of synthetic images. This ideal observer sets an upper bound on the performance of any available observers and provides the best ability to assess whether any differences exist between the distributions of the real and synthetic images.

An ideal observer is defined as a decision strategy that calculates the likelihood ratio of $q_2(\hat{\mathbf{f}})$ and $q_1(\hat{\mathbf{f}})$ and compares the ratio to a threshold. In other words, the ideal observer calculates the test statistic, Λ, given by

$$\Lambda = \frac{q_2(\hat{\mathbf{f}})}{q_1(\hat{\mathbf{f}})}. \tag{2}$$

Our goal is to relate the AUC for this ideal observer to the distance between the distributions of $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$.

Toward this goal, a central component of our derivation is the use of a likelihood-generating function (Barrett *et al* 1998). We first provide the background for the likelihood-generating function in section 2.2. We show that the characteristic functions, which are used to obtain the ideal-observer AUC, can be derived solely based on the likelihood-generating function. Then, in section 2.3, we show that the ideal-observer AUC can be expressed, to an excellent approximation, by the likelihood-generating function evaluated at the origin. More importantly, this value at the origin relates directly to the Bhattacharyya distance between the distributions of the real and the synthetic images. Thus, by using the likelihood-generating function, we are able to establish a direct relationship between the ideal-observer AUC and the similarity in distributions of the real and the synthetic images.

2.2. Background for likelihood-generating function

The likelihood-generating function is central to our derivation as all moments of both Λ and its logarithm, denoted by λ, under hypotheses H_1 and H_2 can be derived. This function was originally introduced by Barrett *et al* (1998), and we follow a similar approach to define the function. Denote the expectation of a random variable t under hypothesis H_j by $\langle t \rangle_j$. We can show that the moments of Λ under H_2 are related to those under H_1 by

$$\langle \Lambda^k \rangle_2 = \int_\infty \mathrm{d}^M\hat{\mathbf{f}}\; q_2(\hat{\mathbf{f}}) \left[\frac{q_2(\hat{\mathbf{f}})}{q_1(\hat{\mathbf{f}})}\right]^k = \int_\infty \mathrm{d}^M\hat{\mathbf{f}}\; q_1(\hat{\mathbf{f}}) \left[\frac{q_2(\hat{\mathbf{f}})}{q_1(\hat{\mathbf{f}})}\right]^{k+1} = \langle \Lambda^{k+1} \rangle_1. \tag{3}$$

Since $\Lambda = \exp(\lambda)$, we can re-write equation (3) as

$$\langle \exp(k\lambda) \rangle_2 = \langle \exp[(k+1)\lambda] \rangle_1. \tag{4}$$

The moment-generating function for a random variable t under hypothesis H_j, denoted by $M_j(\beta)$, is defined by

$$M_j(\beta) = \int_{-\infty}^{\infty} \mathrm{d}t\; \mathrm{pr}(t|H_j)\exp(\beta t) = \langle \exp(\beta t) \rangle_j. \tag{5}$$

Thus, from equation (4), the relationship between the moment-generating functions under the two hypotheses is given by:

$$M_2(\beta) = M_1(\beta + 1). \tag{6}$$

Additionally, the characteristic function for a random variable t under hypothesis H_j, denoted by $\psi_j(\xi)$, is defined by

$$\psi_j(\xi) = \int_{-\infty}^{\infty} \mathrm{d}t\; \mathrm{pr}(t|H_j)\exp(-2\pi \mathrm{i}\xi t). \tag{7}$$

From equations (5) and (7), we readily see that the moment-generating functions and characteristic functions are related to each other by

$$M_j(\beta) = \psi_j\left(\frac{\mathrm{i}\beta}{2\pi}\right). \tag{8}$$

Then, using equations (6) and (8) yields the relationship between the characteristic functions for λ under hypotheses H_1 (class of synthetic images) and H_2 (class of real images):

$$\psi_2(\xi) = \psi_1\left(\xi + \frac{\mathrm{i}}{2\pi}\right). \tag{9}$$

This equation is important since it can further be used to derive the relationship between the probability distributions of λ under the two hypothesis. Denote the probability distribution of λ under hypothesis H_j by $p_j(\lambda)$. Applying inverse Fourier transform to equation (9) on both sides yields (appendix B)

$$p_2(\lambda) = \exp(\lambda) p_1(\lambda). \tag{10}$$

In equation (10), both $p_1(\lambda)$ and $p_2(\lambda)$ can be derived from a single non-negative function $f(\lambda)$, as follows:

$$p_1(\lambda) = \exp\left(-\frac{1}{2}\lambda\right) f(\lambda), \tag{11a}$$

$$p_2(\lambda) = \exp\left(\frac{1}{2}\lambda\right) f(\lambda). \tag{11b}$$

Defining this function $f(\lambda)$ can help us to derive the expressions for the moment-generating functions and characteristic functions now. Denote the two-sided Laplace transform of $f(\lambda)$ by $\mathcal{F}_L(\beta)$, such that

$$\mathcal{F}_L(\beta) = \int_{-\infty}^{\infty} \mathrm{d}\lambda \, \exp(\beta\lambda) f(\lambda). \tag{12}$$

Then, from equation (6), we obtain

$$M_1(\beta) = \mathcal{F}_L\left(\beta - \frac{1}{2}\right), \tag{13a}$$

$$M_2(\beta) = \mathcal{F}_L\left(\beta + \frac{1}{2}\right). \tag{13b}$$

Similarly, $\psi_1(\xi)$ and $\psi_2(\xi)$ in equation (9) can be expressed in terms of the Fourier transform of $f(\lambda)$, denoted by $\mathcal{F}(\xi)$:

$$\psi_1(\xi) = \mathcal{F}\left(\xi - \frac{\mathrm{i}}{4\pi}\right), \tag{14a}$$

$$\psi_2(\xi) = \mathcal{F}\left(\xi + \frac{\mathrm{i}}{4\pi}\right). \tag{14b}$$

The term $p_j(\lambda)$ denotes a probability and should integrate to unity. Thus, from equations (13) and (14), $\mathcal{F}_L\left(\beta \pm \frac{1}{2}\right)$ and $\mathcal{F}\left(\xi \pm \frac{\mathrm{i}}{4\pi}\right)$ should equal to unity. To enforce these constraints, the likelihood-generating function $G(\beta)$ and another function $T(\xi)$ are defined such that

$$\mathcal{F}_L(\beta) = \exp\left[\left(\beta + \frac{1}{2}\right)\left(\beta - \frac{1}{2}\right) G(\beta)\right], \tag{15a}$$

$$\mathcal{F}(\xi) = \exp\left[\left(\xi + \frac{\mathrm{i}}{4\pi}\right)\left(\xi - \frac{\mathrm{i}}{4\pi}\right) T(\xi)\right]. \tag{15b}$$

We can then express $M_1(\beta)$ and $\psi_1(\xi)$ as

$$M_1(\beta) = \exp\left[\beta(\beta - 1) G(\beta - \frac{1}{2})\right], \tag{16a}$$

$$\psi_1(\xi) = \exp\left[\xi(\xi - \frac{\mathrm{i}}{2\pi}) T(\xi - \frac{\mathrm{i}}{4\pi})\right]. \tag{16b}$$

Additionally, from equation (8), $T(\xi)$ can be expressed in terms of $G(\beta)$:

$$T(\xi) = -4\pi^2 G(-2\pi\mathrm{i}\xi). \tag{17}$$

Thus, we see that the characteristic functions can be expressed using only the likelihood-generating function.

2.3. Deriving the relationship between the ideal-observer AUC and the similarity in distributions of the real and the synthetic images

Having obtained the characteristic functions using the likelihood-generating function, we can now derive the expression for the ideal-observer AUC. For this purpose, we note from equation (1) that by expressing the step function in terms of its Fourier transform, we can calculate the AUC as

$$\mathrm{AUC} = \frac{1}{2} + \frac{1}{2\pi\mathrm{i}} \mathscr{P} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\xi} \int_{\infty} \mathrm{d}^M \hat{\mathbf{f}}^s \int_{\infty} \mathrm{d}^M \hat{\mathbf{f}}^r \, q_1(\hat{\mathbf{f}}^s) q_2(\hat{\mathbf{f}}^r) \exp\{2\pi\mathrm{i}\xi[\theta(\hat{\mathbf{f}}^r) - \theta(\hat{\mathbf{f}}^s)]\} \tag{18a}$$

$$= \frac{1}{2} + \frac{1}{2\pi \mathrm{i}} \mathscr{P} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\xi} \left\{ \int_{\infty} \mathrm{d}^M \hat{\mathbf{f}}^s \, q_1(\hat{\mathbf{f}}^s) \exp[-2\pi \mathrm{i}\xi\theta(\hat{\mathbf{f}}^s)] \right\}$$
$$\times \left\{ \int_{\infty} \mathrm{d}^M \hat{\mathbf{f}}^r \, q_2(\hat{\mathbf{f}}^r) \exp[2\pi \mathrm{i}\xi\theta(\hat{\mathbf{f}}^r)] \right\}, \tag{18b}$$

where $\mathscr{P}$ denotes the Cauchy principal value for evaluating the improper integral. Note that in equation (18*b*), the expression within each curly bracket is the same as calculating the expectation of the term $(\pm)2\pi \mathrm{i}\xi\theta(\hat{\mathbf{f}})$. Using the fact that this expectation can be calculated from the probability density on either $\hat{\mathbf{f}}$ or $\theta(\hat{\mathbf{f}})$, we can further write equation (18*b*) in terms of the characteristic functions (equation (7)) as

$$\mathrm{AUC} = \frac{1}{2} + \frac{1}{2\pi \mathrm{i}} \mathscr{P} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\xi} \, \psi_1(\xi) \psi_2^*(\xi). \tag{19}$$

By replacing the expression for $\psi_2(\xi)$ from equation (9) and using the Hermiticity property of the Fourier transform, we obtain

$$\mathrm{AUC} = \frac{1}{2} + \frac{1}{2\pi \mathrm{i}} \mathscr{P} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\xi} \psi_1(\xi) \psi_1\left(-\xi + \frac{\mathrm{i}}{2\pi}\right) \tag{20a}$$

$$= \frac{1}{2} + \frac{1}{2\pi \mathrm{i}} \mathscr{P} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\xi} \exp\left\{ -4\pi^2 \left(\xi^2 - \frac{\mathrm{i}\xi}{2\pi}\right) \left[G(2\pi \mathrm{i}\xi + \frac{1}{2}) + G\left(-2\pi \mathrm{i}\xi - \frac{1}{2}\right) \right] \right\}, \tag{20b}$$

where, in the second step, we have used the expression for $\psi_1(\xi)$ from equation (16*b*) and then the relationship between $T(\xi)$ and $G(\beta)$ from equation (17). To simplify this further, we can approximate $G(\beta)$ via the Maclaurin series expansion:

$$G(\beta) = \sum_{n=0}^{\infty} G^{(n)}(0) \frac{\beta^n}{n!}. \tag{21}$$

Substituting this in equation (20*b*) and assuming that the contribution of higher order ($n > 1$) terms is negligible yields

$$\mathrm{AUC} = \frac{1}{2} + \frac{1}{2\pi \mathrm{i}} \mathscr{P} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\xi} \exp\left\{ -4\pi^2 \left(\xi^2 - \frac{\mathrm{i}\xi}{2\pi}\right) \sum_{n=0}^{\infty} G^{(n)}(0) \frac{\left(2\pi \mathrm{i}\xi + \frac{1}{2}\right)^n + \left(-2\pi \mathrm{i}\xi - \frac{1}{2}\right)^n}{n!} \right\} \tag{22a}$$

$$= \frac{1}{2} + \frac{1}{2\pi \mathrm{i}} \mathscr{P} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\xi} \exp\left\{ -4\pi^2 \left(\xi^2 - \frac{\mathrm{i}\xi}{2\pi}\right) \sum_{k=0}^{\infty} 2G^{(2k)}(0) \frac{\left(2\pi \mathrm{i}\xi + \frac{1}{2}\right)^{2k}}{(2k)!} \right\} \tag{22b}$$

$$\approx \frac{1}{2} + \frac{1}{2\pi \mathrm{i}} \mathscr{P} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\xi} \exp\left\{ -4\pi^2 \left(\xi^2 - \frac{\mathrm{i}\xi}{2\pi}\right) \times 2G(0) \right\}. \tag{22c}$$

By means of tabular integral, equation (22*c*) yields

$$\mathrm{AUC} \approx \frac{1}{2} + \frac{1}{2} \mathrm{erf}\left[\frac{1}{2} \sqrt{2G(0)} \right]. \tag{23}$$

Next, using equations (15*a*), (12), and (11*a*), we obtain

$$G(0) = -4 \log \mathcal{F}_L(0) \tag{24a}$$

$$= -4 \log \int_{-\infty}^{\infty} \mathrm{d}\lambda \, p_1(\lambda) \exp\left(\frac{1}{2}\lambda\right) \tag{24b}$$

$$= -4 \log \langle \Lambda^{\frac{1}{2}} \rangle_1 \tag{24c}$$

$$= -4 \log \left[\int_{\infty} \mathrm{d}^M \hat{\mathbf{f}} \sqrt{q_1(\hat{\mathbf{f}}) q_2(\hat{\mathbf{f}})} \right] \tag{24d}$$

$$= 4 D_B(q_1(\hat{\mathbf{f}}), \, q_2(\hat{\mathbf{f}})), \tag{24e}$$

where, in equation (24*e*), the term $D_B(q_1(\hat{\mathbf{f}}), \, q_2(\hat{\mathbf{f}}))$ is the well-known Bhattacharyya distance (Bhattacharyya 1943) that measures the similarity between the distributions $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$. The term $\int_{\infty} \mathrm{d}^M \hat{\mathbf{f}} \sqrt{q_1(\hat{\mathbf{f}}) q_2(\hat{\mathbf{f}})}$ in equation (24*d*) is the Bhattacharyya coefficient. Then, from equations (23) and (24*e*), we obtain that for an ideal observer, the AUC can be approximated excellently in terms of the Bhattacharyya distance between $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$:

$$\mathrm{AUC} \approx \frac{1}{2} + \frac{1}{2} \mathrm{erf}\left[\sqrt{2 D_B(q_1(\hat{\mathbf{f}}), \, q_2(\hat{\mathbf{f}}))} \right]. \tag{25}$$

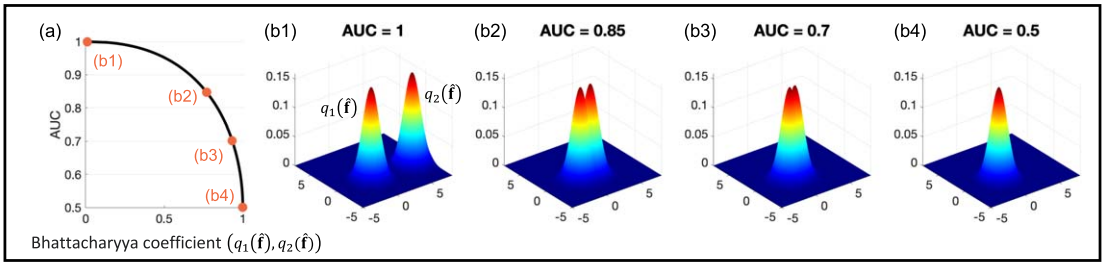

Figure 1. Illustrating the relationship in equation (25) between the ideal-observer AUC and similarity in distributions of $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$ for a two-pixel image setup. (a) The computed AUC values as a function of the Bhattacharyya coefficient between $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$ (equation (24*d*)). (b1-4) The computed AUC values for four representative cases. We note in (b4) that for perfect overlap between $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$, the ideal-observer AUC achieves the lower bound of 0.5.

Note that equation (25) is obtained without making any assumption of the probability law of either the images $\hat{\mathbf{f}}$ or the likelihood ratio Λ.

From equation (25), it is easy to show that the value of the ideal-observer AUC decreases as the Bhattacharyya distance between $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$ decreases, and vice versa. Further, a lower bound of AUC $= 0.5$ is obtained when the Bhattacharyya distance is at the minimum value of 0, i.e. $q_1(\hat{\mathbf{f}})$ exactly matches $q_2(\hat{\mathbf{f}})$. Thus, an ideal-observer-study-based approach provides a mechanism to quantitatively evaluate the similarity in distributions of the real and the synthetic images.

2.4. Illustrating the relationship between the ideal-observer AUC and the Bhattacharyya distance for a two-pixel image setup

To illustrate the relationship in equation (25), consider that $\hat{\mathbf{f}}$ denotes images consisting of only two pixels. For the sake of simplicity, assume that $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$ are described by 2D Gaussian distributions that have the same covariance matrix but different means, i.e. $q_1(\hat{\mathbf{f}}) \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma)$ and $q_2(\hat{\mathbf{f}}) \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma)$. We readily see that the Bhattacharyya distance between $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$ decreases as the difference between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ decreases. Using equation (25), we can obtain the AUC at different values of $D_B(q_1(\hat{\mathbf{f}}),\ q_2(\hat{\mathbf{f}}))$. As shown in figure 1, the value of AUC decreases and achieves the lower bound of 0.5 as the overlap between $q_1(\hat{\mathbf{f}})$ and $q_2(\hat{\mathbf{f}})$ increases, i.e. $D_B(q_1(\hat{\mathbf{f}}),\ q_2(\hat{\mathbf{f}}))$ approaches 0.

3. A web-based expert-human-observer-study-based approach to quantitatively evaluate the clinical realism of synthetic images

As introduced in section 1, human-observer studies have been widely used to evaluate the clinical realism of synthetic images. Among the different human observers, expert human readers, such as physicians who are highly experienced in reading medical images, can identify minute differences between the real and synthetic images. A 2-AFC experiment provides a mechanism to quantify the performance of the expert human observers on this task. If an expert human observer correctly identifies the real images for only around 50% of the cases in the 2-AFC experiment, then, as mentioned in section 2.1 with the proof provided in appendix A, this would indicate an AUC of ∼0.5 on the task of detecting the real image. This would imply that the expert human observer was unable to distinguish between the real and synthetic images, thus, suggesting that the synthetic images are clinically realistic as evaluated by that observer.

While several tools have been developed for conducting human-observer studies (Håkansson *et al* 2010, Zhang *et al* 2016), users often need to manually install the tools on local workstations with compatible operating systems and/or have programming knowledge. These requirements can reduce the accessibility of the tools and consequently, serve as a hurdle in designing and conducting the observer studies. To address these issues, we develop an openly available software for conducting the 2-AFC experiments by expert human observers to quantitatively evaluate the clinical realism of synthetic images. This software is designed to be accessible, secure, and have mechanisms for both designing new 2-AFC experiments by investigators and performing the experiments by expert human observers. To achieve these goals, we design this software to be web-based and with a dual-user 'Investigator-Reader' interface. The 'Investigator interface' allows an investigator to design a 2-AFC experiment and upload the real and the synthetic images. The 'Reader interface' allows the expert human observers recruited by this investigator to perform the 2-AFC experiment. The programming environment for

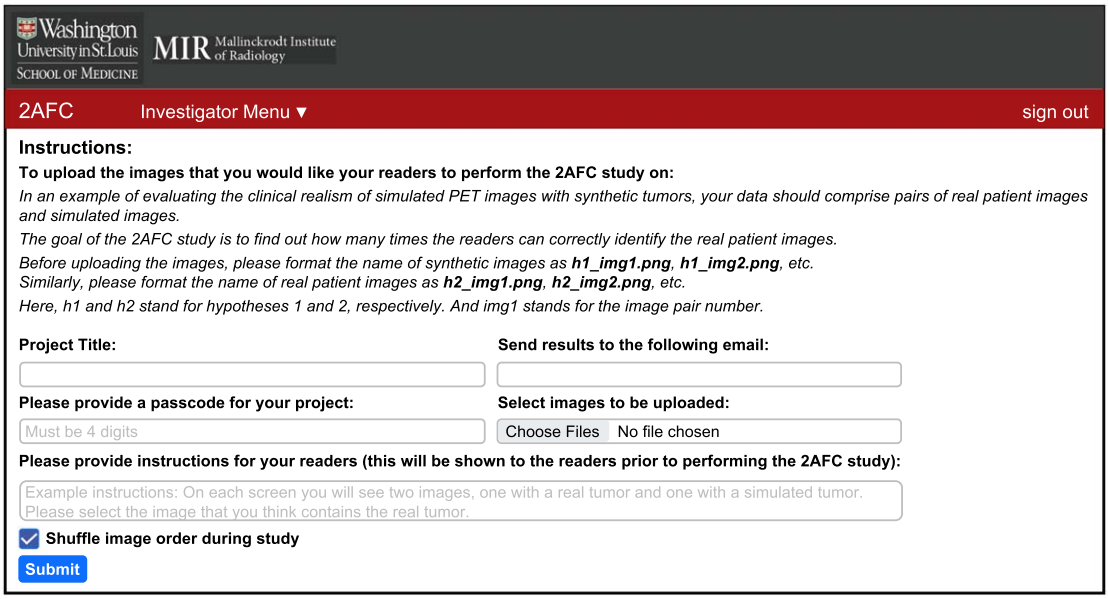

Figure 2. The investigator interface.

building the software is detailed in appendix C. In the following, we focus on describing the main functionalities of this software and the procedures for the investigator and reader to design and perform the 2-AFC experiment.

3.1. Developed software

3.1.1. Investigator interface

The layout for the investigator interface is shown in figure 2. As a first step, the investigator is required to provide a project title and a corresponding four-digit passcode, which the investigator should then share with the readers. This ensures that only readers authorized by this investigator can access the images, thus ensuring the security of the images. To improve the accessibility for readers, the investigator is asked to provide instructions for the readers to perform the 2-AFC experiment on the uploaded images. These instructions will be displayed on the screen once a reader begins the experiment. Our software allows the investigator to upload an arbitrary number of image pairs. The investigator is also provided an option to shuffle the order of image pairs. Finally, the investigator is asked to provide an email address, to which the results of the observer study from each reader would be sent. Note that if an investigator receives results with a percent accuracy much lower than 50%, this is likely an indication that the observer is not trained and, thus, the results should be treated with caution.

3.1.2. Reader interface

The reader is required to provide the project title and the corresponding passcode to access the images uploaded by a specific investigator. If these entries are provided correctly, the reader will be directed to the webpage, as shown in figure 3, to perform the 2-AFC experiment. In this experiment, a synthetic image sampled from $q_1(\hat{\mathbf{f}})$ and a real image sampled from $q_2(\hat{\mathbf{f}})$ are presented side-by-side (section 2.1). For each image pair, the reader is asked to identify the image that they perceive as real. While making the decision, the reader can adjust the contrast and invert the intensities of the images. The goal of providing these functionalities is to increase the clinical relevance and rigor of the observer study. The reader is also asked to provide a confidence level for the decision. The interpretations of the confidence levels are provided to the reader (figure 3). These interpretations are similar to those used in previous studies to conduct human-observer studies (Chen *et al* 2016, Ma *et al* 2017). The confidence levels could be a useful tool for improving the design of the synthesis technique after the observer-study evaluation. For example, if an expert reader correctly distinguishes the real image from the synthetic image with high confidence level, this could indicate that the synthetic image is highly unrealistic. Investigators could then incorporate such feedback while improving the design of their synthetic-image-generation approaches. Additionally, the reader is provided with an option to leave additional comments.

3.2. Evaluating usability of the developed software

To evaluate the usability of our software, we conduct a system usability scale (SUS) survey (Brooke 1996). This survey is widely used to test the usability of newly developed software and websites. The SUS evaluates a software on three main aspects, namely, effectiveness, efficiency, and satisfaction. These aspects assess whether users

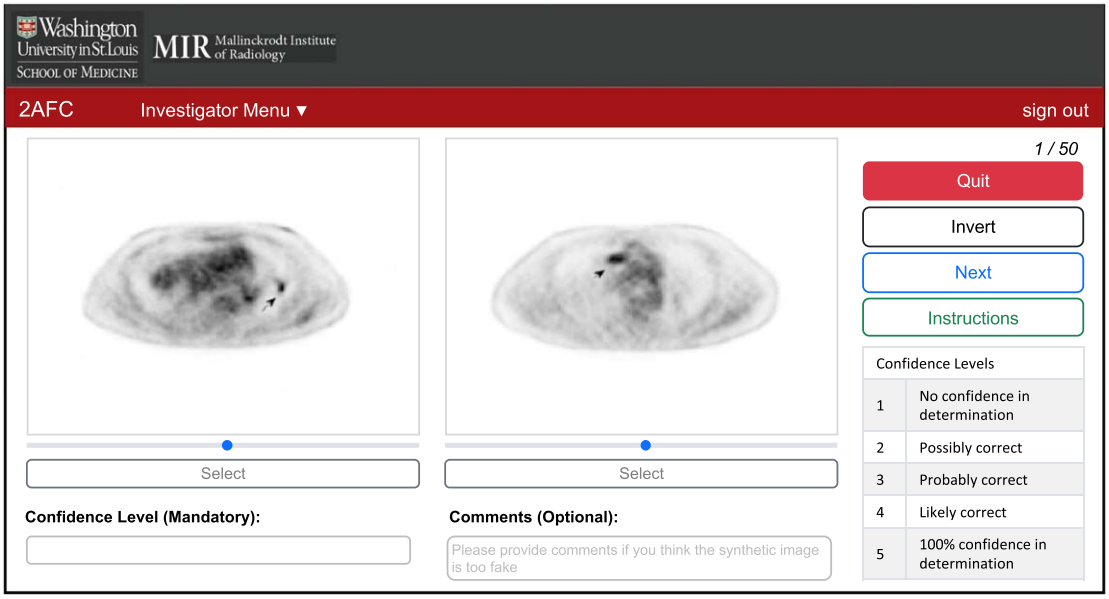

Figure 3. The reader interface.

Table 1. The system usability scale (SUS) survey.

Index	Statement
1	I think that I would like to use this software frequently.
2	I found the software unnecessarily complex.
3	I thought the software was easy to use.
4	I think that I would need the support of a technical person to be able to use this software.
5	I found the various functionalities of this software were well integrated.
6	I thought there was too much inconsistency in this software.
7	I would imagine that most people would learn to use this software very quickly.
8	I found the software very cumbersome to use.
9	I felt very confident using the software.
10	I needed to learn a lot of things before I could get going with this software.

achieve their goals successfully, the effort and/or resource spent to achieve the goals, and whether the user experience is satisfactory, respectively.

The SUS survey was designed by adapting from Brooke (1996) and consisted of a 10-item questionnaire about the software with five response options for respondents: strongly disagree, disagree, neutral, agree, and strongly agree (table 1). For the odd-numbered items, a score of 0 was assigned to 'strongly disagree' and a score of 4 was assigned to 'strongly agree'. For the even-numbered items, a score of 4 was assigned to 'strongly disagree' and a score of 0 was assigned to 'strongly agree'. The scores were then added, and the summed score was multiplied by 2.5 such that the eventual score fell between 0 and 100.

We first conducted the survey with five board-certified nuclear medicine physicians with years of expertise ranging from 7 to 40 years (median: 12 years, average: 20.4 years), one nuclear medicine physicist, and one nuclear medicine resident. These users are considered as the expert human observers who would use our software to evaluate the clinical realism of synthetic images. Additionally, we conducted the survey with five users who were asked to evaluate the software as investigators designing an observer study. Conducting the survey with all these users provides evidence for the utility of the software in practical settings.

3.3. Evaluating the clinical realism of a positron emission tomography (PET) image-synthesis technique using the developed software

To demonstrate the application of our software to quantitatively evaluate the clinical realism of image-synthesis techniques, we used the software to evaluate a recently developed technique for oncologic PET. This technique is a stochastic and physics-based method that generates 2D ^{18}F-fluorodeoxyglucose (FDG)-PET images of patients with lung cancer (Liu *et al* 2021a). By following the simulation procedure detailed in Liu *et al* (2021a), we

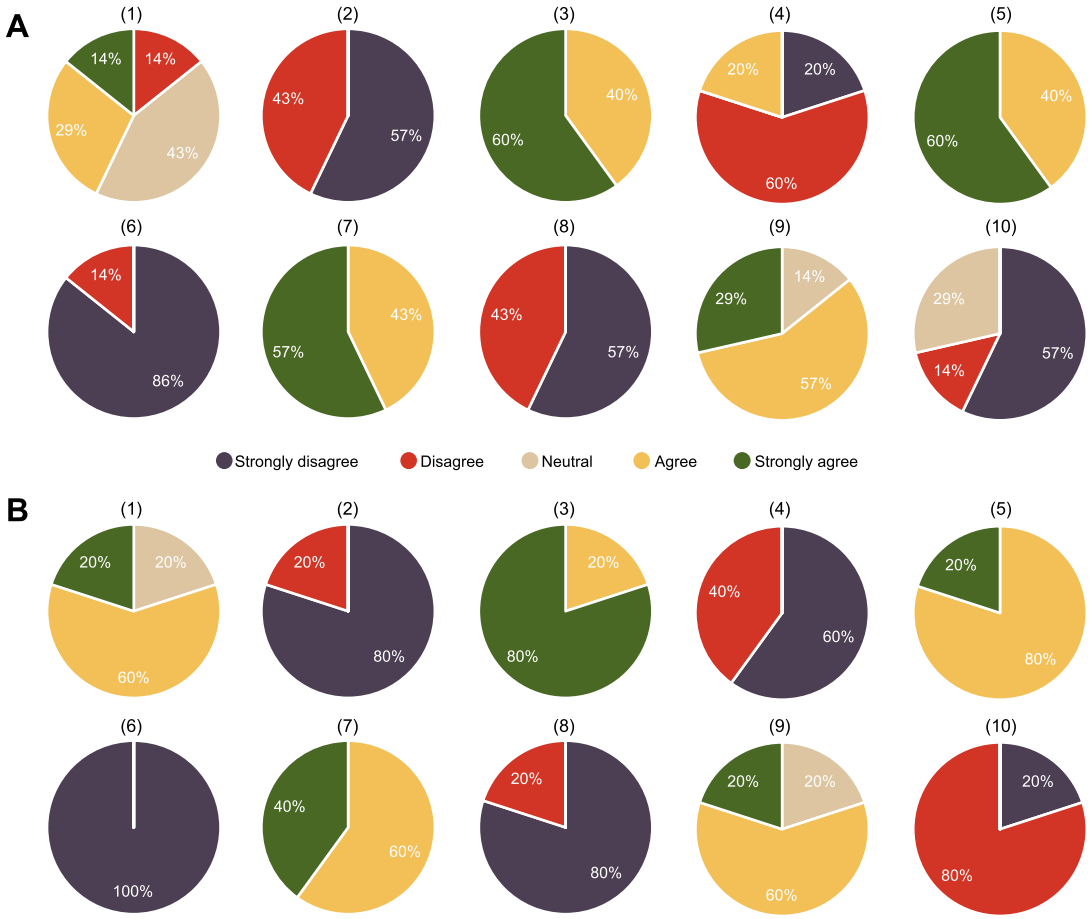

Figure 4. Distribution of responses to each item in the questionnaire from (A) seven expert human readers and (B) five observer-study designers participating in the system usability scale (SUS) survey.

generated 50 synthetic PET images for our 2-AFC study. The source code for this technique is openly available at https://github.com/ziping-liu/A-stochastic-and-physics-based-method-to-generate-oncological-PET-images.git. Our evaluation study was retrospective, involved clinical images, and was IRB-approved and HIPAA-compliant with informed consent being waived.

The 2-AFC study using our developed software was conducted by six expert readers, including five board-certified PET physicians (BAS, FD, JCM, TJF, and MI) and one PET physicist (RL). The readers were highly experienced in reading PET scans, with years of expertise ranging from 7 to 40 years (median: 16 years, average: 20.3 years). During the study, each of the 50 synthetic images was paired with an existing clinical PET image to be displayed to the readers simultaneously with our software (section 3.1.2; figure 3). The readers were then asked to identify the real image, provide a confidence level for the decision, and optionally leave a comment. We then computed the percentage of times that each reader correctly identified the real PET image.

4. Results

4.1. Evaluating usability of the developed software for conducting 2-AFC experiments with expert human observers

In this section, we report the outcome of the SUS survey conducted to evaluate the usability of the developed web application (section 3.2). Figure 4 presents the distribution of responses from (A) seven expert human readers and (B) five observer-study designers to each item in the questionnaire described in table 1. Figure 5 shows the total score computed for each user based on the rule defined in section 3.2. For the group of expert human readers, a mean score of 84 with standard deviation of 8 was observed. Similarly, a mean score of 87 with standard deviation of 5 was obtained for the group of investigators. Based on Lewis and Sauro (2018), these results indicate that our software is very highly usable.

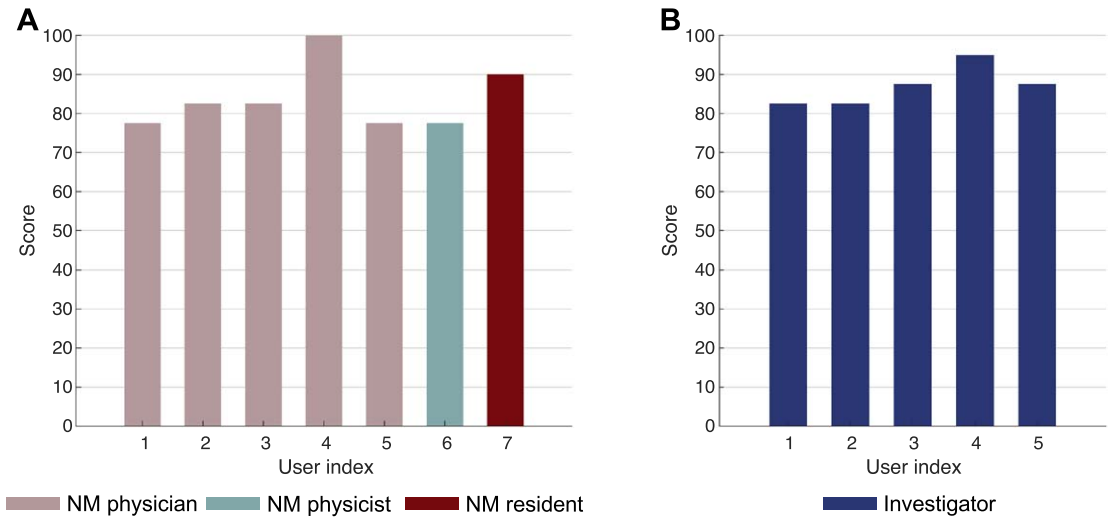

Figure 5. Total score for each user participating in the system usability scale (SUS) survey. (NM: nuclear medicine).

Table 2. Percent accuracy and median confidence level for each expert reader participating in the 2-AFC study.

Reader	Percent accuracy	Median confidence level
PET physician 1	44%	2
PET physician 2	58%	4
PET physician 3	50%	2
PET physician 4	58%	3
PET physician 5	44%	4
PET physicist	58%	4

4.2. Evaluating the clinical realism of a PET image-synthesis technique using the developed software

Table 2 shows the percent accuracy and median confidence level for each expert human observer participating in the 2-AFC study to evaluate the clinical realism of the stochastic and physics-based image-synthesis technique using our developed software, as described in section 3.3. We observe that all the readers identified the real PET image correctly only ~50% of the time. Additionally, for half of the readers, the median value of confidence levels was $\leqslant 3$.

Figure 6 shows the number of correct (upper row) and incorrect (lower row) decisions made by the (a) five PET physicians, (b) the PET physicist, and (c) all the readers, respectively, at each confidence level. When combining all the readers, only 164/300 (55%) decisions were made correctly. Among these correct decisions, only 71 (43%) were made with confidence levels $\geqslant 4$. Additionally, 34/136 (25%) incorrect decisions were made with high confidence levels $\geqslant 4$.

5. Discussion

To ensure that simulation-based development and evaluation of medical imaging methods are clinically relevant, images generated by the synthesis technique must be clinically realistic and, ideally, have the same distribution as that of real images. The first contribution of this work is to theoretically demonstrate that an ideal-observer-study-based approach provides a mechanism to quantitatively evaluate the similarity in distributions between the real and synthetic images. Further, we show that the AUC for an ideal observer can be expressed, to an excellent approximation, by the Bhattacharyya distance between the distributions of real and synthetic images. Thus, when the ideal-observer AUC decreases, this indicates that the distance between the two distributions decreases. Moreover, a lower bound of AUC $= 0.5$ indicates that the distribution of the synthetic images exactly matches that of the real images. Thus, by quantifying the similarity in distributions between the real and synthetic images, this ideal-observer-study-based approach provides a theoretical foundation for quantitative evaluation of the clinical realism of synthetic images.

The second contribution of this manuscript is to develop a web-based platform for facilitating the use of human-observer-study-based approaches to quantitatively evaluate the clinical realism of synthetic images. Our

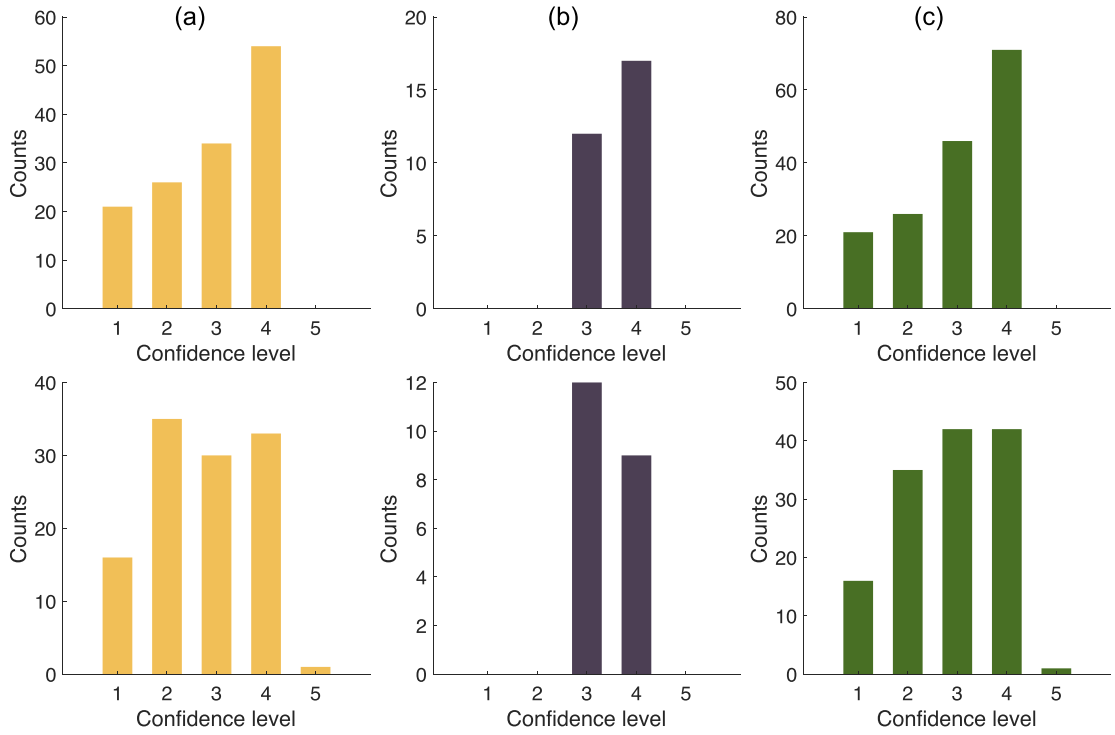

Figure 6. Number of correct (upper row) and incorrect (lower row) decisions made by the (a) five PET physicians, (b) one physicist, and (c) all the readers, at each confidence level.

software is openly available, does not require installation on a local workstation, is platform-independent, eliminates the need for on-site study, and allows simultaneous access by multiple users. The goal of incorporating all these features is to strengthen the usability of this software. Additionally, our software provides features that allow varying the contrast and intensity of images. This leads to an user interface that is similar to those present in clinical tools, thus further strengthening the rigor and clinical relevance of the 2-AFC experiments. Our results from the SUS survey shown in section 4.1 demonstrate that the software is highly user-friendly and accessible. Further, our software provides multiple features to align with the General Data Protection Regulation policies. Specifically, the software provides mechanisms to secure stored data, allow users to delete uploaded data, and prevent data from unauthorized access. All these features are important for evaluation studies that include patient data.

Our developed software can be used to evaluate a large class of image-synthesis techniques, including physics-based methods (Duchateau *et al* 2017, Ma *et al* 2017, Leung *et al* 2020, Hamdi *et al* 2021), generative adversarial network-based methods (Costa *et al* 2017, Nie *et al* 2017, Wang *et al* 2021), and other AI-based methods (Chartsias *et al* 2017b, Xiang *et al* 2018, Bahrami *et al* 2020, Dutta *et al* 2022). Further, while the key purpose of our software is evaluating the realism of synthetic images, the software can also be used to conduct 2-AFC experiments for performing image-quality assessment. For this secondary purpose, tools have been developed previously (Vuong *et al* 2018, Genske and Jahnke 2022). Similar to those tools, our software can be used to evaluate newly developed image-reconstruction and image-processing methods on signal-detection tasks.

Another application of the proposed realism-evaluation strategies is in assessing the realism of synthetic images that are generated for virtual clinical trials. For this application, it is important to account for the clinical task of interest and not just assess whether the images look realistic to a human observer (Badano 2017). In that context, our ideal-observer-study-based approach provides a mechanism to quantify the difference in distributions of real and synthetic images. Further, performance on clinical tasks of interest typically depends on the distribution of the image. Future research may reveal that having a measure of the difference between the distributions of real and synthetic images can help to objectively compare the performance on the clinical task with those images. In that case, our theoretical formalism could provide a mechanism to account for the clinical task of interest when evaluating the realism of synthetic images.

As a secondary finding of this work, our evaluation of a stochastic and physics-based image-synthesis technique (section 3.3) using the expert-human-observer-based study with the developed software indicates that the expert readers had limited ability to distinguish the real images from the synthetic images. As shown in table 2, all the expert readers, even including the most experienced PET physician with 40 years of reading PET

scans, correctly identified the real images only in ∼50% of the cases. Additionally, we observe from figure 6 that among the 164 (out of 300) correct decisions, only 43% were made with high confidence levels, suggesting that the readers were not confident even when they correctly identified the real image. Moreover, the readers were falsely confident for 25% of incorrect decisions. These results motivate the use of the image-synthesis technique to generate images for the development and evaluation of a wide range of PET imaging methods. In fact, this technique was used to objectively evaluate a recently developed PET segmentation method (Liu *et al* 2021b).

There are some limitations in this work. First, our ideal-observer-study-based approach to evaluate the clinical realism of synthetic images was presented in theory and not yet applied to a clinical scenario. As shown in section 2, developing the ideal observer requires knowledge of the probability distributions of the real and synthetic images. However, in clinical studies, these distributions are high-dimensional and do not have a known analytical form. To address these issues, AI-based methods are showing promise in approximating the ideal-observer test statistics for signal-detection tasks (Kupinski *et al* 2001, Zhou *et al* 2019b). Our theoretical formalism motivates extending these methods for the task of clinical realism evaluation. Second, our theoretical formalism was presented specifically for an ideal observer and thus, we reiterate that it should not be used to directly interpret results obtained with expert human observers. However, in that context, we do point out that several studies (He *et al* 2004, Li *et al* 2016) have shown correlations between the performance of human observers and channelized Hotelling observers (CHOs). The CHOs utilize templates that are derived from the first- and second-order statistics of the channel vectors extracted from the images. Thus, in special cases where the channel vectors are sufficient statistics for describing the distributions of real and synthetic images, our ideal-observer analysis may be used to quantify the similarity in distributions of real and synthetic images. Examining this connection is an important future research direction. A third limitation is that our web application is currently designed to evaluate the realism of synthetic images on a per-slice basis and not the entire 3D volume. Additionally, in the designed application, the slices are displayed only in a single orientation. Expanding the web application to display images in 3D and in multiple orientations is an important area of future development. Finally, our web application is currently developed for conducting 2-AFC experiments. Considering that different variants of the 2-AFC experiment have been used in the human-observer studies (Zhang *et al* 2016, Ikejimba *et al* 2019), expanding our software to allow conducting those experiments is another important area of future development.

6. Conclusion

In this work, we investigated two observer-study-based approaches to quantitatively evaluate the clinical realism of synthetic images. We theoretically demonstrated that an ideal-observer-study-based approach provides a mechanism to quantify the similarity in distributions of real and synthetic images. Further, we showed that the ideal-observer AUC can be expressed, to an excellent approximation, by the Bhattacharyya distance between the distributions of real and synthetic images. Additionally, we developed a software that provides a web-based platform to facilitate the conducting of expert-human-observer studies for quantitative evaluation of the realism of synthetic images. This software is available at https://apps.mir.wustl.edu/twoafc. The software provides multiple functionalities towards increasing the rigor and clinical relevance of 2-AFC experiments. Our results from the SUS survey demonstrate that this software enables designing and performing 2-AFC experiments with expert human observers in a highly accessible and user-friendly manner. Finally, as a secondary finding of this work, evaluation of a stochastic and physics-based PET image-synthesis technique showed that the expert human observers were generally unable to distinguish the real images from the synthetic images. This finding motivates the application of this technique to the development and evaluation of PET imaging methods.

Acknowledgments

Financial support for this work was provided by the National Institute of Biomedical Imaging and Bioengineering R01-EB031051, R56-EB028287 and R01-EB031962. We also thank Qiye Tan for the help with initial development of web application for conducting the observer study.

Ethical statement

The study was conducted under an approved IRB protocol titled 'PET Radiomics of NSCLC' (ID: 201906021) at Washington University in St. Louis. We confirm that the research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements. This was a retrospective study and waiver of consent was granted.

Appendix A

In this appendix, we prove that when an observer performs a 2-AFC experiment, the expression for the probability of a correct decision (equation (1)) is equal to the AUC for that observer. Our proof is similar to that provided in Barrett *et al* (1998) but for a different context. In that paper, the derivation was presented in the context of performing a 2-AFC study to evaluate the observer performance for a signal-detection task. Here, we paraphrase the derivation for the application of evaluating the clinical realism of synthetic images.

Proof. Consider an observer performing the task of identifying an image as synthetic (H_1) or real (H_2). For a given image, the observer calculates a test statistic, denoted by a random variable t, and then compares the value of t to a threshold, denoted by x. If $t \geqslant x$, the observer will identify the image as real, i.e. assign the image to H_2. Otherwise, the image is considered synthetic and assigned to H_1.

The performance of this observer can be fully specified by two quantities. The first quantity, referred to as the true-positive fraction (TPF), measures the fraction of times that the observer identifies the image as real when the image is indeed real. The second quantity, referred to as false-positive fraction (FPF), measures the fraction of times that the observer identifies the image as real when the image is in fact synthetic. Denote the probability of an event by $\Pr(\cdot)$ and the probability distribution of a random variable by $\mathrm{pr}(\cdot)$. Given the threshold x, the TPF and FPF can be calculated as follows:

$$\mathrm{TPF}(x) = \Pr(t \geqslant x|H_2) = \int_x^\infty \mathrm{d}t\ \mathrm{pr}(t|H_2), \tag{A.1a}$$

$$\mathrm{FPF}(x) = \Pr(t \geqslant x|H_1) = \int_x^\infty \mathrm{d}t\ \mathrm{pr}(t|H_1). \tag{A.1b}$$

Then, the expression for the AUC can be obtained in terms of the TPF and FPF as

$$\mathrm{AUC} = \int_0^1 \mathrm{dFPF}(x)\ \mathrm{TPF}(x) \tag{A.2a}$$

$$= -\int_{-\infty}^\infty \mathrm{d}x\ \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{FPF}(x)\mathrm{TPF}(x), \tag{A.2b}$$

where, in the second step, we have changed the variable of integration from $\mathrm{FPF}(x)$ to x. For convenience of notation, we define $p_j(t) \equiv \mathrm{pr}(t|H_j)$. Using equation (A.1b) and Leibniz's rule, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{FPF}(x) = -p_1(x). \tag{A.3}$$

Then, we can re-write equation (A.2b) as

$$\mathrm{AUC} = \int_{-\infty}^\infty \mathrm{d}x\ p_1(x) \int_x^\infty \mathrm{d}t\ p_2(t) \tag{A.4a}$$

$$= \int_{-\infty}^\infty \mathrm{d}x\ \int_{-\infty}^\infty \mathrm{d}t\ p_1(x)p_2(t)\mathrm{step}(t-x). \tag{A.4b}$$

Since the test statistic is calculated based on the image $\hat{\mathbf{f}}$, we consider t as a function of $\hat{\mathbf{f}}$ such that $t = \theta(\hat{\mathbf{f}})$. Thus, we can further write the AUC in equation (A.4) in terms of integrals over $\hat{\mathbf{f}}$. For this purpose, we first express the term $p_j(t)$ as

$$p_j(t) = \int_\infty \mathrm{d}^M\hat{\mathbf{f}}\ \mathrm{pr}(t|\hat{\mathbf{f}})\mathrm{pr}(\hat{\mathbf{f}}|H_j) \tag{A.5a}$$

$$= \int_\infty \mathrm{d}^M\hat{\mathbf{f}}\ \mathrm{pr}(\hat{\mathbf{f}}|H_j)\delta[t-\theta(\hat{\mathbf{f}})], \tag{A.5b}$$

where, in the second step, the function $t = \theta(\hat{\mathbf{f}})$ is represented as a probabilistic mapping. For convenience of notation, we define $q_j(\hat{\mathbf{f}}) \equiv \mathrm{pr}(\hat{\mathbf{f}}|H_j)$. Then, from equations (A.4b) and (A.5), we obtain

$$\mathrm{AUC} = \int_{-\infty}^\infty \mathrm{d}x\ \int_{-\infty}^\infty \mathrm{d}t\ \int_\infty \mathrm{d}^M\hat{\mathbf{f}}\ q_1(\hat{\mathbf{f}})\delta[x-\theta(\hat{\mathbf{f}})] \int_\infty \mathrm{d}^M\hat{\mathbf{f}}'\ q_2(\hat{\mathbf{f}}')\delta[t-\theta(\hat{\mathbf{f}}')]\mathrm{step}(t-x) \tag{A.6a}$$

$$= \int_\infty \mathrm{d}^M\hat{\mathbf{f}}\ q_1(\hat{\mathbf{f}}) \int_\infty \mathrm{d}^M\hat{\mathbf{f}}'\ q_2(\hat{\mathbf{f}}') \int_{-\infty}^\infty \mathrm{d}x\ \delta[x-\theta(\hat{\mathbf{f}})] \int_{-\infty}^\infty \mathrm{d}t\ \delta[t-\theta(\hat{\mathbf{f}}')]\mathrm{step}(t-x) \tag{A.6b}$$

$$= \int_\infty \mathrm{d}^M\hat{\mathbf{f}} \int_\infty \mathrm{d}^M\hat{\mathbf{f}}'\ q_1(\hat{\mathbf{f}})q_2(\hat{\mathbf{f}}')\mathrm{step}(\theta(\hat{\mathbf{f}}')-\theta(\hat{\mathbf{f}})), \tag{A.6c}$$

where in the second step, we have interchanged the order of integration and, in the third step, we have used the sifting property of the delta function to perform the integrals over x and t. We then immediately see that equation (A.6c) has the same form as in equation (1). □

Appendix B

In this appendix, we provide the derivation for obtaining the relationship between the probability distribution of the log-likelihood ratio λ under the two hypothesis (equation (10)).

We first apply inverse Fourier transform to equation (9) on both sides:

$$p_2(\lambda) = \int_{-\infty}^{\infty} \mathrm{d}\xi \ \psi_1\left(\xi + \frac{\mathrm{i}}{2\pi}\right)\exp(2\pi \mathrm{i}\xi\lambda). \tag{B.1}$$

By letting $z = \xi + \frac{\mathrm{i}}{2\pi}$, we have

$$p_2(\lambda) = \exp(\lambda)\int_{-\infty+\frac{\mathrm{i}}{2\pi}}^{\infty+\frac{\mathrm{i}}{2\pi}} \mathrm{d}z \ \psi_1(z)\exp(2\pi \mathrm{i}z\lambda). \tag{B.2}$$

As proven in Barrett *et al* (1998), the contour in equation (B.2) can be shifted as long as $\langle\Lambda\rangle_2$ is finite, thus, yielding

$$p_2(\lambda) = \exp(\lambda)\int_{-\infty}^{\infty} \mathrm{d}z \ \psi_1(z)\exp(2\pi \mathrm{i}z\lambda) \tag{B.3a}$$

$$= \exp(\lambda)p_1(\lambda). \tag{B.3b}$$

Appendix C

In this appendix, we describe the programming environment of the developed web application.

The web application was developed on Microsoft's .NET 6 software framework and leveraged the Razor Pages web development model. The Razor Pages model incorporates the model-view-viewmodel design pattern to facilitate separation of the user-interface layer from the backend domain layer. The model-view-viewmodel pattern is an object-oriented-programming paradigm characterized by the use of an intermediary viewmodel object that serves to expose data within model objects for presentation to the user within the view. The application is primarily written in the C# programming language for server-side operations, and adopts the object-oriented-programming approach. The application also consists of a client-side layer written in JavaScript to deliver responsive and dynamic user-interface functionalities to the user. The client-side layer integrates data retrieved from the server with the user-interface to create functionality that is not dependent on additional HTTP requests to render. The Module Pattern is used to scope client-side code to defined areas within the application and encapsulate application logic. Application data is persisted within a SQL Server relational database instance which communicates with the application through Microsoft's Entity Framework object-relational-mapping tool. A Microsoft Azure B2C tenant instance was employed to handle user access and authentication into the application. The B2C tenant also handles third-party authentication for the application.

ORCID iDs

Ziping Liu https://orcid.org/0000-0001-8107-6522
Zitong Yu https://orcid.org/0000-0002-5917-1918
Abhinav K Jha https://orcid.org/0000-0002-4795-4878

References

Abadi E, Segars W P, Tsui B M W, Kinahan P E, Bottenus N, Frangi A F, Maidment A, Lo J and Samei E 2020 Virtual clinical trials in medical imaging: a review *J. Med. Imaging* **7** 042805

Badano A 2017 How much realism is needed?—the wrong question in silico imagers have been asking *Med. Phys.* **44** 1607–9

Badano A 2021 In silico imaging clinical trials: cheaper, faster, better, safer, and more scalable *Trials* **22** 1–7

Badano A, Graff C G, Badal A, Sharma D, Zeng R, Samuelson F W, Glick S J and Myers K J 2018 Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an in silico imaging trial *JAMA Network Open* **1** e185474e185474–185474

Bahrami A, Karimian A, Fatemizadeh E, Arabi H and Zaidi H 2020 A new deep convolutional neural network design with efficient learning capability: application to CT image synthesis from MRI *Med. Phys.* **47** 5158–71

Barrett H H, Abbey C K and Clarkson E 1998 Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions *J. Opt. Soc. Am.* A **15** 1520–35

Barrett H H and Myers K J 2013 *Foundations of Image Science* (New York: Wiley)

Bhattacharyya A 1943 On a measure of divergence between two statistical populations defined by their probability distributions *Bull. Calcutta Math. Soc.* **35** 99–109

Brooke J 1996 SUS-A quick and dirty usability scale *In Usability Evaluation in Industry* (Boca Raton, FL: CRC Press)

Burgess A E 2011 Visual perception studies and observer models in medical imaging *Semin. Nucl. Med.* **41** 419–36

Chartsias A, Joyce T, Dharmakumar R and Tsaftaris S A 2017a Adversarial image synthesis for unpaired multi-modal cardiac data *International Workshop on Simulation and Synthesis in Medical Imaging* (Berlin: Springer) pp 3–13

Chartsias A, Joyce T, Giuffrida M V and Tsaftaris S A 2017b Multimodal MR synthesis via modality-invariant latent representation *IEEE Trans. Med. Imaging* **37** 803–14
Chen B, Ma C, Leng S, Fidler J L, Sheedy S P, McCollough C H, Fletcher J G and Yu L 2016 Validation of a projection-domain insertion of liver lesions into CT Images *Acad. Radiol.* **23** 1221–9
Costa P, Galdran A, Meyer M I, Niemeijer M, Abràmoff M, Mendonça A M and Campilho A 2017 End-to-end adversarial retinal image synthesis *IEEE Trans. Med. Imaging* **37** 781–91
Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B and Bharath A A 2018 Generative adversarial networks: an overview *IEEE Signal Process. Mag.* **35** 53–65
Duchateau N, Sermesant M, Delingette H and Ayache N 2017 Model-based generation of large databases of cardiac images: synthesis of pathological cine MR sequences from real healthy cases *IEEE Trans. Med. Imaging* **37** 755–66
Dutta K, Liu Z, Laforest R, Jha A and Shoghi K I 2022 Deep learning framework to synthesize high-count preclinical PET images from low-count preclinical PET images *Medical Imaging 2022: Physics of Medical Imaging* (San Diego, California, United States: SPIE) vol 12031, pp 351–60
Elangovan P, Mackenzie A, Dance D R, Young K C, Cooke V, Wilkinson L, Given-Wilson R M, Wallis M G and Wells K 2017 Design and validation of realistic breast models for use in multiple alternative forced choice virtual clinical trials *Phys. Med. Biol.* **62** 2778–94
Frangi A F, Tsaftaris S A and Prince J L 2018 Simulation and synthesis in medical imaging *IEEE Trans. Med. Imaging* **37** 673–9
Genske U and Jahnke P 2022 Human observer net: a platform tool for human observer studies of image data *Radiology* **303** 524–30
Gong K, Guan J, Liu C-C and Qi J 2018 PET image denoising using a deep neural network through fine tuning *IEEE Trans. Radiat. Plasma Med. Sci.* **3** 153–61
Gonias P *et al* 2007 Validation of a GATE model for the simulation of the Siemens biograph^{TM} 6 PET scanner *Nucl. Instrum. Methods Phys. Res.* A **571** 263–6
Guan S and Loew M 2019 Using generative adversarial networks and transfer learning for breast cancer detection by convolutional neural networks *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications* (San Diego, California, United States: SPIE) vol 10954, pp 306–18
Håkansson M, Svensson S, Zachrisson S, Svalkvist A, BÅth M and MÅnsson L G 2010 VIEWDEX: an efficient and easy-to-use software for observer performance studies *Radiat. Prot. Dosim.* **139** 42–51
Hamdi M, Natsuaki Y, Wangerin K A, An H, James S S, Kinahan P E, Sunderland J J, Larson P E Z, Hope T A and Laforest R 2021 Evaluation of attenuation correction in PET/MRI with synthetic lesion insertion *J. Med. Imaging* **8** 056001
He X, Frey E C, Links J M, Gilland K L, Segars W P and Tsui B M W 2004 A mathematical observer study for the evaluation and optimization of compensation methods for myocardial SPECT using a phantom population that realistically models patient variability *IEEE Trans. Nucl. Sci.* **51** 218–24
Hernandez-Giron I, den Harder J M, Streekstra G J, Geleijns J and Veldkamp W J H 2019 Development of a 3D printed anthropomorphic lung phantom for image quality assessment in CT *Phys. Med.* **57** 47–57
Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S 2017 GANs trained by a two time-scale update rule converge to a local nash equilibrium *Adv. Neural. Inf. Process. Syst.* **30** 6629–40
Houbrechts K, Vancoillie L, Cockmartin L, Marshall N W and Bosmans H 2021 Virtual clinical trial platforms for digital breast tomosynthesis: a local solution compared to the VICTRE platform *Medical Imaging 2021: Physics of Medical Imaging* (Online Only: SPIE) vol 11595, 403–11
Ikejimba L C, Salad J, Graff C G, Ghammraoui B, Cheng W-C, Lo J Y and Glick S J 2019 A four-alternative forced choice (4AFC) methodology for evaluating microcalcification detection in clinical full-field digital mammography (FFDM) and digital breast tomosynthesis (DBT) systems using an inkjet-printed anthropomorphic phantom *Med. Phys.* **46** 3883–92
Jha A K, Caffo B and Frey E C 2016 A no-gold-standard technique for objective assessment of quantitative nuclear-medicine imaging methods *Phys. Med. Biol.* **61** 2780–800
Jha A K, Myers K J, Obuchowski N A, Liu Z, Rahman M A, Saboury B, Rahmim A and Siegel B A 2021 Objective task-based evaluation of artificial intelligence-based medical imaging methods: framework, strategies, and role of the physician *PET Clin.* **16** 493–511
Jha A K *et al* 2022 Nuclear medicine and artificial intelligence: best practices for evaluation (the RELAINCE guidelines) *J. Nucl. Med.* **63** 1288–99
Kupinski M A, Edwards D C, Giger M L and Metz C E 2001 Ideal observer approximation using Bayesian classification neural networks *IEEE Trans. Med. Imaging* **20** 886–99
Leung K H, Marashdeh W, Wray R, Ashrafinia S, Pomper M G, Rahmim A and Jha A K 2020 A physics-guided modular deep-learning based automated framework for tumor segmentation in PET *Phys. Med. Biol.* **65** 245032
Lewis J R and Sauro J 2018 Item benchmarks for the system usability scale *J. Usability Stud.* **13** (3) 158–67
Li X, Jha A K, Ghaly M, Elshahaby F E A, Links J M and Frey E C 2016 Use of sub-ensembles and multi-template observers to evaluate detection task performance for data that are not multivariate normal *IEEE Trans. Med. Imaging* **36** 917–29
Li Z, Benabdallah N, Abou D S, Baumann B C, Dehdashti F, Jammalamadaka U, Laforest R, Wahl R L, Thorek D L J and Jha A K 2022 A projection-domain low-count quantitative SPECT method for α-particle emitting radiopharmaceutical therapy *IEEE Trans. Radiat. Plasma Med. Sci.* **7** 62–74
Liu Z, Laforest R, Mhlanga J, Fraum T J, Itani M, Dehdashti F, Siegel B A and Jha A K 2021a Observer study-based evaluation of a stochastic and physics-based method to generate oncological PET images *Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment* (Online Only: SPIE) vol 11599, pp 9–17
Liu Z, Mhlanga J C, Laforest R, Derenoncourt P-R, Siegel B A and Jha A K 2021b A Bayesian approach to tissue-fraction estimation for oncological PET segmentation *Phys. Med. Biol.* **66** 124002
Liu Z, Moon H S, Li Z, Laforest R, Perlmutter J S, Norris S A and Jha A K 2022 A tissue-fraction estimation-based segmentation method for quantitative dopamine transporter SPECT. *Med. Phys.* **49** (8) 5121–37
Ma C *et al* 2017 Evaluation of a projection-domain lung nodule insertion technique in thoracic Computed Tomography *J. Med. Imaging* **4** 013510
Maidment A D A 2014 Virtual clinical trials for the assessment of novel breast screening modalities *Int. Workshop on Digital Mammography* (Berlin: Springer) pp 1–8
Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q and Shen D 2017 Medical image synthesis with context-aware generative adversarial networks *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer) pp 417–25
Poon J K, Dahlbom M L, Casey M E, Qi J, Cherry S R and Badawi R D 2015 Validation of the SimSET simulation package for modeling the Siemens Biograph mCT PET scanner *Phys. Med. Biol.* **60** N35–45

Shmelkov K, Schmid C and Alahari K 2018 How good is my GAN? *Proceedings of the European Conference on Computer Vision (ECCV)* pp 213–29

Song N, Du Y, He B and Frey E C 2011 Development and evaluation of a model-based downscatter compensation method for quantitative I-131 SPECT *Med. Phys.* **38** 3193–204

Sturgeon G M, Park S, Segars W P and Lo J Y 2017 Synthetic breast phantoms from patient based eigenbreasts *Med. Phys.* **44** 6270–9

Surti S, Karp S, Popescu L M, Daube-Witherspoon E and Werner M 2006 Investigation of time-of-flight benefit for fully 3-DPET *IEEE Trans. Med. Imaging* **25** 529–38

Vuong J, Kaur S, Heinrich J, Ho B K, Hammang C J, Baldi B F and O'Donoghue S I 2018 VersusA tool for evaluating visualizations and image quality using a 2AFC methodology *Vis. Inform.* **2** 225–34

Wang C, Yang G, Papanastasiou G, Tsaftaris S A, Newby D E, Gray C, Macnaught G and MacGillivray T J 2021 DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis *Inf. Fusion* **67** 147–60

Xiang L, Wang Q, Nie D, Zhang L, Jin X, Qiao Y and Shen D 2018 Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image *Med. Image Anal.* **47** 31–44

Yousefirizi F, Jha A K, Brosch-Lenz J, Saboury B and Rahmim A 2021 Toward high-Throughput artificial intelligence-based segmentation in oncological PET imaging *PET Clin.* **16** 577–96

Yu Z, Rahman M A, Schindler T, Gropler R, Laforest R, Wahl R and Jha A 2020 AI-based methods for nuclear-medicine imaging: need for objective task-specific evaluation *J. Nucl. Med.* **61** 575–

Zhang G, Cockmartin L and Bosmans H 2016 A four-alternative forced choice (4AFC) software for observer performance evaluation in radiology *Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment* (San Diego, California, United States: SPIE) vol 9787, pp 369–74

Zhou W, Bhadra S, Brooks F and Anastasio M A 2019a Learning stochastic object model from noisy imaging measurements using AmbientGANs *Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment* (San Diego, California, United States: SPIE) vol 10952, pp 142–8

Zhou W, Li H and Anastasio M A 2019b Approximating the Ideal Observer and Hotelling observer for binary signal detection tasks by use of supervised learning methods *IEEE Trans. Med. Imaging* **38** 2456–68