

AAPM Task Group Report 273: Recommendations on Best Practices for AI and Machine Learning for Computer-Aided Diagnosis in Medical Imaging

*Original*

AAPM Task Group Report 273: Recommendations on Best Practices for AI and Machine Learning for Computer-Aided Diagnosis in Medical Imaging / Hadjiiski, Lubomir; Cha, Kenny; Chan, Heang-Ping; Drukker, Karen; Morra, Lia; Näppi, Janne J.; Sahiner, Berkman; Yoshida, Hiroyuki; Chen, Quan; Deserno, Thomas M.; Greenspan, Hayit; Huisman, Henkjan; Huo, Zhimin; Mazurchuk, Richard; Petrick, Nicholas; Regge, Daniele; Samala, Ravi; Summers, Ronald M.; Suzuki, Kenji; Tourassi, Georgia; Vergara, Daniel; Armato III, Samuel G.. - In: MEDICAL PHYSICS. - ISSN 0094-2405. - 50:2(2023), pp. 1-24. [10.1002/mp.16188]

*Availability:*

This version is available at: 11583/2973414 since: 2023-01-06T15:19:18Z

*Publisher:*

Wiley

*Published*

DOI:10.1002/mp.16188

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

Wiley postprint/Author's Accepted Manuscript

This is the peer reviewed version of the above quoted article, which has been published in final form at <http://dx.doi.org/10.1002/mp.16188>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

(Article begins on next page)

1 **AAPM Task Group Report 273: Recommendations on Best**  
2 **Practices for AI and Machine Learning for Computer-Aided**  
3 **Diagnosis in Medical Imaging**

4  
5 Lubomir Hadjiiski

6 *Department of Radiology, University of Michigan, Ann Arbor, Michigan, USA*  
7

8 Kenny Cha

9 *U.S. Food and Drug Administration, Silver Spring, Maryland, USA*  
10

11 Heang-Ping Chan

12 *Department of Radiology, University of Michigan, Ann Arbor, Michigan, USA*  
13

14 Karen Drukker

15 *Department of Radiology, University of Chicago, Chicago, Illinois, USA*  
16

17 Lia Morra

18 *Department of Control and Computer Engineering, Politecnico di Torino, Torino, Italy*  
19

20 Janne J. Näppi

21 *3D Imaging Research, Department of Radiology, Massachusetts General Hospital and Harvard*  
22 *Medical School, Boston, Massachusetts, USA*  
23

24 Berkman Sahiner

25 *U.S. Food and Drug Administration, Silver Spring, Maryland, USA*  
26

27 Hiroyuki Yoshida

28 *3D Imaging Research, Department of Radiology, Massachusetts General Hospital and Harvard*  
29 *Medical School, Boston, Massachusetts, USA*  
30

31 Quan Chen

32 *Department of Radiation Medicine, University of Kentucky, Lexington, Kentucky, USA*  
33

34 Thomas M. Deserno

35 *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical*  
36 *School, Braunschweig, Germany*  
37

38 Hayit Greenspan

39 *Department of Biomedical Engineering, Faculty of Engineering, Tel Aviv University, Tel Aviv,*  
40 *Israel & Department of Radiology, Ichan School of Medicine, Mt Sinai, NYC, NY, USA*  
41  
42

43 Henkjan Huisman  
44 *Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The*  
45 *Netherlands*  
46  
47 Zhimin Huo  
48 *Tencent America, Palo Alto, CA*  
49  
50 Richard Mazurchuk  
51 *Division of Cancer Prevention, National Cancer Institute, National Institutes of Health,*  
52 *Bethesda, Maryland, USA*  
53  
54 Nicholas Petrick  
55 *U.S. Food and Drug Administration, Silver Spring, Maryland, USA*  
56  
57 Daniele Regge  
58 *Radiology Unit, Candiolo Cancer Institute, FPO-IRCCS, Candiolo*  
59 *Department of Surgical Sciences, University of Turin, Turin, Italy*  
60  
61 Ravi Samala  
62 *U.S. Food and Drug Administration, Silver Spring, Maryland, USA*  
63  
64 Ronald M. Summers  
65 *Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda,*  
66 *Maryland, USA*  
67  
68 Kenji Suzuki  
69 *Institute of Innovative Research, Tokyo Institute of Technology, Tokyo, Japan*  
70  
71 Georgia Tourassi  
72 *Oak Ridge National Lab, Oak Ridge, Tennessee, USA*  
73  
74 Daniel Vergara  
75 *Department of Radiology, Yale New Haven Hospital, New Haven, Connecticut, USA*  
76  
77 Samuel G. Armato, III  
78 *Department of Radiology, University of Chicago, Chicago, Illinois, USA*  
79  
80

## 81 **Disclosure Statement**

82 The Chair of the AAPM Task Group 273 has reviewed the required Conflict of Interest statement  
83 on file for each member of AAPM Task Group 273 and determined that disclosure of potential  
84 Conflicts of Interest is an adequate management plan. Disclosures of potential Conflicts of  
85 Interest for each member of AAPM Task Group 273 are found at the close of this document.  
86

87	<b>Table of Contents</b>
88	
89	<b>Abstract</b>
90	
91	<b>1 Introduction</b>
92	
93	<b>2 Data</b>
94	<b>2.1 Data Collection</b>
95	<b>2.1.1 Data collection and case sampling</b>
96	<b>2.1.2 Public databases</b>
97	<b>2.1.3 Ethics considerations of data collection</b>
98	<b>2.1.3.1 De-identification</b>
99	<b>2.1.3.2 Diversity and Inclusion</b>
100	<b>2.1.4 Quality considerations</b>
101	<b>2.2 Data Augmentation</b>
102	<b>2.3 Data Harmonization</b>
103	<b>2.4 Take Home Message on Data</b>
104	
105	<b>3 Reference Standards</b>
106	<b>3.1 Objective vs. Subjective Reference Standards</b>
107	<b>3.2 Annotation Granularity</b>
108	<b>3.2.1 Entire image</b>
109	<b>3.2.2 Region-based</b>
110	<b>3.2.3 Pixel-based</b>
111	<b>3.3 Methods for Acquiring Annotations</b>
112	<b>3.3.1 Expert labels</b>
113	<b>3.3.2 Electronic health record</b>
114	<b>3.3.3 Crowd sourcing</b>
115	<b>3.3.4 Phantoms</b>
116	<b>3.3.5 Weak/noisy labels</b>
117	<b>3.4 Definition of True Positives</b>
118	<b>3.5 Take Home Message on Reference Standards</b>
119	
120	<b>4 Model Development</b>
121	<b>4.1 Data Sampling Strategies</b>
122	<b>4.2 Machine Learning Strategies</b>
123	<b>4.2.1 Levels of learning supervision</b>
124	<b>4.2.1.1 Supervised learning</b>
125	<b>4.2.1.2 Semi-supervised learning</b>
126	<b>4.2.1.3 Self-supervised learning</b>
127	<b>4.2.1.4 Unsupervised learning</b>
128	<b>4.2.1.5 Multiple-instance learning</b>
129	<b>4.2.2 Transfer learning, multi-task learning, and domain adaptation</b>
130	<b>4.2.2.1 Transfer learning</b>
131	<b>4.2.2.2 Multi-task learning</b>
132	<b>4.2.2.3 Domain adaptation</b>

133	<b>4.2.3 Federated learning</b>
134	<b>4.2.4 “Continuous learning” systems</b>
135	<b>4.3 Take Home Message on Model Development</b>
136	
137	<b>5 Performance Assessment</b>
138	<b>5.1 Performance Assessment Metrics</b>
139	<b>5.2 Statistical Significance</b>
140	<b>5.3 Intended Use</b>
141	<b>5.4 Standalone Performance Assessment</b>
142	<b>5.5 Clinical Reader Performance Assessment</b>
143	<b>5.6 Sample Size</b>
144	<b>5.7 Reproducibility</b>
145	<b>5.8 Take Home Message on Performance Evaluation</b>
146	
147	<b>6 Translation to Clinic</b>
148	<b>6.1 Human-Machine Interface</b>
149	<b>6.2 User Training</b>
150	<b>6.3 Acceptance Testing</b>
151	<b>6.4 Prospective Surveillance</b>
152	<b>6.4.1 Periodic quality assurance</b>
153	<b>6.4.2 Performance monitoring for “continuous learning” systems</b>
154	<b>6.4.3 Prospective evaluation of CAD-AI</b>
155	<b>6.5 Take Home Message on Translation to Clinic</b>
156	
157	<b>7 Discussion</b>
158	
159	<b>Conclusions</b>
160	
161	<b>Disclosure Statement</b>
162	
163	<b>Acknowledgments</b>
164	
165	<b>References</b>
166	

167 **Report of AAPM Task Group 273**

168  
169 The purpose of this report is to provide recommendations on best practices and standards for the  
170 development and performance assessment of computer-aided decision support systems at the time  
171 when machine learning techniques continue to evolve, and CAD applications expand to new  
172 stages of the patient care process. The various steps of development are covered, including (1)  
173 data collection, (2) establishing reference standards, (3) model development, (4) performance  
174 assessment, and (5) translation to clinical practice. The goal of the report is to emphasize the  
175 proper training and validation methods for machine learning algorithms that may improve their  
176 generalizability and reliability and accelerate the adoption of CAD-AI systems for clinical  
177 decision support.

178  
179

180 **Abstract**

181 Rapid advances in artificial intelligence (AI) and machine learning, and specifically in deep  
182 learning (DL) techniques, have enabled broad application of these methods in health care. The  
183 promise of the DL approach has spurred further interest in computer-aided diagnosis (CAD)  
184 development and applications using both ‘traditional’ machine learning methods and newer DL-  
185 based methods. We use the term CAD-AI to refer to this expanded clinical decision support  
186 environment that uses traditional and DL-based AI methods.

187 Numerous studies have been published to date on the development of machine learning tools  
188 for computer-aided, or AI-assisted, clinical tasks. However, most of these machine learning  
189 models are not ready for clinical deployment. It is of paramount importance to ensure that a  
190 clinical decision support tool undergoes proper training and rigorous validation of its  
191 generalizability and robustness before adoption for patient care in the clinic.

192 To address these important issues, the American Association of Physicists in Medicine  
193 (AAPM) Computer-Aided Image Analysis Subcommittee (CADSC) is charged, in part, to  
194 develop recommendations on practices and standards for the development and performance  
195 assessment of computer-aided decision support systems. The committee has previously  
196 published two opinion papers on the evaluation of CAD systems and issues associated with user  
197 training and quality assurance of these systems in the clinic. With machine learning techniques  
198 continuing to evolve and CAD applications expanding to new stages of the patient care process,  
199 the current task group report considers the broader issues common to the development of most, if  
200 not all, CAD-AI applications and their translation from the bench to the clinic. The goal is to  
201 bring attention to the proper training and validation of machine learning algorithms that may  
202 improve their generalizability and reliability and accelerate the adoption of CAD-AI systems for  
203 clinical decision support.

204  
205

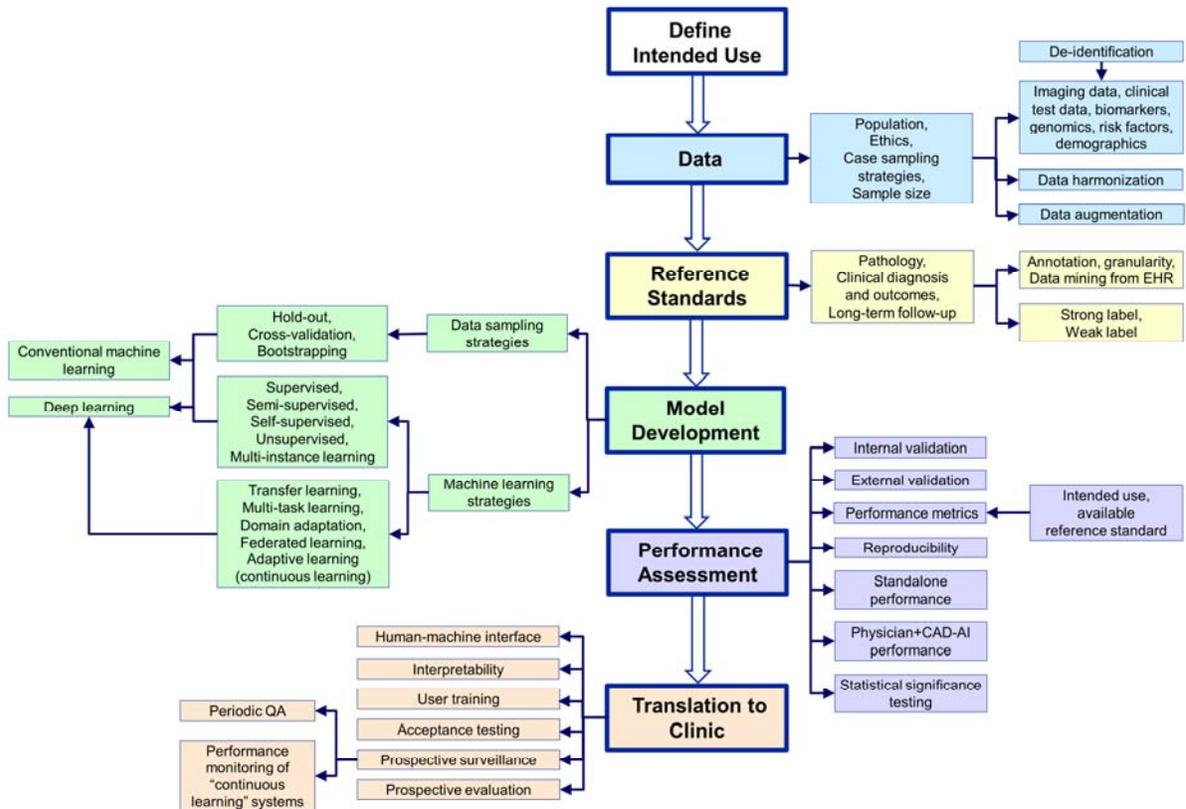
206 **1 Introduction**

207  
208 We are witnessing extensive development and an explosion of applications based on deep  
209 learning (DL) or “artificial intelligence (AI)” technology across various fields in recent years.  
210 Many applications in robotics, transportation, surveillance, Internet, and popular games have  
211 achieved high degrees of success and raised unprecedented enthusiasm for AI. Rapid advances in



245 recommended for robust validation of the real-world clinical performance of AI algorithms.  
 246 These reviews reveal that the majority of machine learning models developed to date seem to be  
 247 far from ready for clinical deployment despite the reported levels of performance.

248 Regardless of the underlying machine learning methods used for development of CAD tools,  
 249 it is of paramount importance to ensure that a clinical decision support tool has undergone proper  
 250 training and rigorous validation of its generalizability and robustness before the adoption of such  
 251 tools for patient care in the clinic. To address these important issues, the American Association  
 252 of Physicists in Medicine (AAPM) Computer-Aided Image Analysis Subcommittee (CADSC) is  
 253 charged, in part, to develop recommendations on practices and standards for the development and  
 254 performance assessment of computer-aided decision support systems. The CADSC has  
 255 previously published two papers to convey the opinions of CADSC members on proper practices  
 256 for the training, evaluation, and quality assurance of CAD systems [5, 6]. With machine learning  
 257 techniques continuing to evolve and CAD applications expanding to new stages of the patient  
 258 care process (Figure 1), this task group report addresses the broad issues common to the  
 259 development of most, if not all, CAD-AI applications and their translation from the bench to the  
 260 clinic. The various steps of development will be covered, including data collection, establishing  
 261 reference standards, model development, performance assessment, and translation to clinical  
 262 practice, as summarized in Figure 2. The goal is to bring attention to proper training and  
 263 validation methods for machine learning algorithms that may improve their generalizability and  
 264 reliability and accelerate the adoption of CAD-AI systems for clinical decision support.  
 265



266  
 267  
 268  
 269  
 270

Figure 2. Overview of development of computer-aided decision support systems

271

## 272 **2 Data**

273

274 The most fundamental step for the development of a CAD-AI tool is to define the use case  
275 and the population to which the CAD-AI tool is to be applied. As a guiding principle, data  
276 collected for the training, validation, and testing of a CAD-AI tool should reflect the intended use  
277 case and population while at the same time allowing for the replication of results in a real-world  
278 clinical setting. It cannot be overemphasized that improper data collection practices may likely  
279 introduce bias and create a misleading perception of model performance, especially in  
280 subpopulations that may not be appropriately represented in the study dataset. In study reports,  
281 the data collection process must be described in detail to demonstrate scientific rigor and should  
282 include inclusion and exclusion criteria as well as the target patient demographics.

283 This section covers the topics of data collection (including case sampling, public databases,  
284 ethics, and quality considerations), data augmentation, and data harmonization. The topic of  
285 labels that might accompany collected data will be covered in the Reference Standards section  
286 (section 3).

287

### 288 **2.1 Data Collection**

289

#### 290 **2.1.1 Data collection and case sampling**

291 System development with consecutively sampled cases from multiple sites over a defined  
292 period of image acquisition dates [7] is the best way to achieve replication of performance in a  
293 real-world clinical setting. In some machine learning applications for which the proportion of  
294 different case groups is highly imbalanced in the population, however, consecutive data  
295 collection is impractical, and the training dataset must be collected with methods such as  
296 stratified sampling to enrich some of the groups. For example, in the case of screening  
297 mammography, stratifying samples across the positive and negative groups is needed because the  
298 yield of malignancy is only 0.5%. **Stratified sampling** [8] splits the population into non-  
299 overlapping groups (or strata) and then samples within each strata to achieve the desired balance  
300 among different strata; if applied accurately, stratified sampling can enhance the generalizability  
301 of a model relative to training without stratification. In practice, many development studies are  
302 performed using a **convenience sample approach** [9], where cases that are conveniently  
303 available to the developers are the ones collected for the study. Especially in new research areas,  
304 the availability of only a convenience sample should not prevent a study from going forward;  
305 however, claims about CAD-AI system performance in such studies should be made with utmost  
306 care to reflect the reality that the results are likely not generalizable.

307 Several recent studies have indicated that systems developed and tested with data from one  
308 collection site failed to achieve similar test results when applied to data from a different site [10-  
309 13]. For this reason, especially for validation studies, it is essential to have **multi-site data**  
310 **collection** [14, 15] and to assure that the data collection is diverse in terms of subject population,  
311 disease severity, vendor/imaging system, and image acquisition protocol. Development studies  
312 that use **single-site data collection** are essential for new advancements in a time-efficient  
313 manner, but strong limitations about the assessed performance should be acknowledged.

314

#### 315 **2.1.2 Public databases**

316 In CAD-AI development, each research group typically uses its resources to collect its own  
317 database, which is likely to be smaller in number than desirable and lacking the real-world

318 diversity of patient demographics and image acquisition parameters that exist across institutions.  
319 Furthermore, this isolation of databases prohibits the direct comparison of the performance of  
320 systems reported in the literature [16, 17].

321 **Publicly available image databases** overcome these shortcomings by providing a free,  
322 accessible resource for the international medical imaging research community. The creation of a  
323 **public database** is not as simple as depositing one or more existing local databases on a web site  
324 or crowd-sourcing the uploading of images and associated information. The nature of the public  
325 database should be prospectively determined in terms of the clinical task(s) it may be expected to  
326 address, the range of disease presentations to be represented by those cases, the associated  
327 metadata it will include, and the reference standard it will incorporate. The need for a quality  
328 assurance (QA) process for data in a public database cannot be overemphasized [18, 19]:  
329 adherence to the case inclusion/exclusion criteria, proper de-identification of protected health  
330 information (PHI), image quality, and reference standard integrity must all be verified before the  
331 database can be released for public access. In addition, the FAIR (Findable, Accessible,  
332 Interoperable and Reusable) principles must be followed to the extent possible in designing  
333 public datasets to assist both human users and their computational agents in the discovery of,  
334 access to, and integration and analysis of the data [20].

335 Public databases are resources of growing importance for the advancement of machine  
336 learning algorithms in medical imaging and clinical decision support in general. These databases  
337 play important roles in algorithm development, training/testing, validation, and performance  
338 assessment; in short, they expedite the ability of research groups to contribute to the field.  
339 Investigators who use these databases have an obligation to understand the limitations of the  
340 databases and to use them in a manner consistent with the capabilities they offer.

341

### 342 **2.1.3 Ethics considerations of data collection**

343 The rapid advancement of machine learning in medicine has prompted new questions about  
344 the **legal framework and ethics of data collection**. The **legal framework** varies by country. In  
345 the United States, the Health and Human Services (HHS) Privacy Rule standards [21] address the  
346 use and disclosure of individuals' PHI, which includes information in a medical record that can  
347 be linked to a specific individual. For research, the Privacy Rule stipulates that covered entities  
348 are permitted to use and disclose PHI (1) with individual authorization or (2) without individual  
349 authorization under "limited circumstances" that must be approved by Institutional Review Board  
350 (IRB). In the European Union, the General Data Protection Regulation (GDPR) provides the  
351 framework for data protection and includes considerations for the use of healthcare data for a  
352 purpose different from the one for which it was originally collected (secondary use) with and  
353 without explicit patient consent. Many other countries have also established guidelines or  
354 regulations on ethics considerations for the use of human subject data [22]. For example, China  
355 released Personal Information Security Specification in 2018 to promote privacy rules established  
356 in their 2017 Cyber Security Law as a national standard [23, 24]. Brazil established the Brazilian  
357 General Data Protection Law (LGPD) in 2020; while it is broadly aligned with the EU GDPR,  
358 some notable differences exist [25]. Independent of legal considerations, several authors have  
359 recently argued for an ethical framework in which the secondary use of clinical data without  
360 explicit patient consent is ethically justifiable, as long as mechanisms are in place to ensure that  
361 ethical standards are strictly followed [26]. Additional issues related to **ethics of data collection**  
362 for machine learning systems in medical imaging include: (1) de-identification of PHI in medical  
363 images and other supporting data, and (2) impact of data collection on algorithm fairness [27].

364

365

366

### 367 **2.1.3.1 De-identification**

368 **De-identification** refers to removal or encoding of identifiers from patient health information  
369 collected for research purposes. In radiological imaging, many of these identifiers are present in  
370 the DICOM header contained within each image file when the image is generated for patient care  
371 purposes, and several toolkits offer a number of different strategies for de-identification of  
372 DICOM headers. For example, the Radiologic Society of North America’s Clinical Trials  
373 Processor is a tool that is recommended for de-identifying DICOM headers when optimal  
374 security is required, due to its high level of customization [28]. De-identification of DICOM  
375 headers, however, may be insufficient for some radiological datasets, because there may exist  
376 potential sources of PHI other than those within the DICOM header [29]: actual pixels within the  
377 image (“burned-in” data) might contain PHI, especially in ultrasound images and radiographs;  
378 objects worn by a patient that contain personal information (such as a bracelet) may appear in  
379 medical images; and data in head-and-neck CT images may allow facial reconstruction that could  
380 identify the patient. For these reasons, it is advisable to visually inspect images and use additional  
381 tools for optimal security, especially if the images are to be publicly shared.

382

### 383 **2.1.3.2 Diversity and Inclusion**

384 A potentially significant, yet subtle, consequence of improper data collection might be an  
385 algorithm that performs poorly for certain subgroups or subpopulations with the targeted disease  
386 or condition as a result of under-representation of those subgroups in the training set [30, 31]. In  
387 radiology applications, it is important to be vigilant so that training/validation dataset selection  
388 incorporates safeguards to minimize underlying distortions for under-represented and/or  
389 vulnerable populations and so that already-existing health-care inequities are not perpetuated or  
390 exacerbated [27, 32-34].

391

### 392 **2.1.4 Quality considerations**

393 **Image quality** may have a strong impact on the reported performance of CAD-AI systems.  
394 Fortunately, many imaging centers have an image QA program already in place, and imaging  
395 exams are typically repeated if the image quality is substandard. Nevertheless, it is still good  
396 practice to ensure that a QA program is being followed at image collection sites and to visually  
397 inspect key images to ensure image quality is acceptable before entering a case into a database  
398 for CAD-AI training, if feasible.

399 An additional consideration is whether the images were acquired with equipment that is still  
400 technically relevant and in accordance with appropriate image acquisition protocols. This ensures  
401 that a CAD-AI system trained or tested with the dataset is capable of answering clinically  
402 relevant questions. With rapid advances in image acquisition hardware and software, a collected  
403 dataset can quickly become obsolete. To create an enduring image dataset, data collection and  
404 management should be considered a continuous process rather than a one-shot effort.

405 Consideration of data curation is essential to the integrity of an image dataset. The dataset  
406 should be inspected (either visually or by automated analysis) to ensure that it contains only  
407 images from the relevant anatomic site and image modality. It is important to be aware of the  
408 differences in image acquisition parameters, imaging time points, selected series from CT scans,  
409 contrast enhancement status, and contrast administration timing. A more subtle point for data  
410 curation involves awareness of the potential bias that may be introduced if “positive” cases, for  
411 example, come from one site or scanner while all “negative” cases come from a different site or  
412 scanner, a situation that should be avoided. If developing a multi-institutional dataset, curation

413 should be performed at the institutional level, where local clinical information is more easily  
414 accessible and verifiable, before depositing to the dataset, if possible.

415

## 416 **2.2 Data Augmentation**

417 **Data augmentation** is a collection of task-dependent techniques used to create alterations of  
418 the training data or to create synthetic data to increase the training set size aiming to improve the  
419 generalization that may be achieved by a trained CAD-AI algorithm [35]. Data augmentation has  
420 become an essential part of the training process for CAD-AI algorithms due to the recent use of  
421 deep neural networks that have millions of parameters and thus require a large number of training  
422 iterations for adequate training. To create variations of existing images contained within the  
423 training set, early successful deep learning applications for image classification used  
424 parameterized transformations that included affine transformations such as image rotation,  
425 flipping, scaling, and jittering [36]. Non-rigid transformations such as deformable  
426 transformations were later used for data augmentation.

427 Data augmentation based on the recently developed technique of generative adversarial  
428 networks [37] has attracted strong interest. Generative adversarial neural networks have the  
429 ability to learn the underlying data distribution and to generate synthetic images mimicking the  
430 actual ones that may fill the gaps in feature distributions [38]. Other approaches to data  
431 augmentation include obtaining images from physical phantoms or generating synthetic data from  
432 physics modeling [39]. Physical and virtual phantoms have been used in medical imaging for  
433 development of new imaging techniques, improvement of existing imaging modalities, and the  
434 conduct of virtual clinical trials; images generated from these approaches represent a natural  
435 extension for data augmentation.

436 Data augmentation techniques that create alterations of the training data should not modify  
437 the image appearance in a manner that makes the underlying biological or tissue properties  
438 implausible. In addition, it should be recognized that these techniques can only generate slight  
439 variations to the structural properties of existing samples in the training set; they cannot create  
440 new patterns or independent information that do not exist in the original training set. Although  
441 data augmentation may help the machine learning algorithm better interpolate among existing  
442 samples, it cannot fundamentally compensate for the inadequacies of a small clinical training set.  
443 The use of synthetic data (in silico and phantom) may prove useful for creating large training sets  
444 if the real-world variabilities of the clinical task, and the human subjects, and the imaging system  
445 can be realistically modeled. It remains to be shown that these synthetic data can sufficiently  
446 simulate the physiological or biological properties of real patients required for developing  
447 decision support tools for many clinical tasks.

448

## 449 **2.3 Data Harmonization**

450 Data may include images obtained at different sites, acquired with different equipment and  
451 image-acquisition parameters, and reconstructed and/or post-processed using different  
452 algorithms. These differences may result in systematic variations across images. **Data**  
453 **harmonization** aims to reduce these variations retrospectively after acquisition while preserving  
454 the biological variability captured in the images [40]. Technically, DL-based methods are capable  
455 of handling variations in image appearance provided the training dataset includes example cases  
456 capturing all those variations and each in sufficient number to provide adequate learning;  
457 however, the demands of such inclusion on dataset collection and subsequent training could  
458 become prohibitively resource intensive. Moreover, deep learning methods can learn which site  
459 an image came from (for multi-institutional datasets) or which vendor's equipment was used for  
460 image acquisition, so utmost care should be taken to minimize bias in the training data [11]. For

461 example, if all mammograms with breast cancer were acquired on a mammography unit from  
462 vendor A and all mammograms with benign lesions were acquired on a mammography unit from  
463 vendor B, a deep learning method is apt to learn to distinguish images from vendor A from those  
464 from vendor B rather than to distinguish the salient imaging features between breast cancers and  
465 benign lesions.

466 In practice, data harmonization has become the key to enhancing accuracy and robustness of  
467 CAD-AI systems [36, 41]. Researchers should be aware of the heterogeneity of image appearance  
468 and quality (and record, for example, differences in image acquisition parameters) during the data  
469 collection stage and incorporate data harmonization methods, when appropriate, to aid models in  
470 accommodating data heterogeneity [42, 43]. Harmonization methods can be applied in the image  
471 domain or feature-space domain [44]. Image-domain harmonization methods include post-  
472 processing of image data [45] and style transfer [46], and feature-domain harmonization methods  
473 include basic statistical normalization techniques [47] and advanced statistical techniques such as  
474 ComBat [48, 49]. The Quantitative Imaging Biomarkers Alliance (QIBA) and the Quantitative  
475 Imaging Network (QIN) have also devoted efforts to the harmonization of medical imaging data  
476 and tools [50, 51]. It is important to recognize that although data harmonization aims to reduce  
477 the systematic variations due to image acquisition, reconstruction, and post-processing or due to  
478 different protocols among data collection sites, it does not address the issue of systematic  
479 variations among patient sub-populations (see sections 2.1.3.2 and 4.2.2.3).

480

## 481 **2.4 Take Home Message on Data**

482 In summary, proper data collection methods are of critical importance to successful training,  
483 validation, and implementation of CAD-AI algorithms. Improper collection and manipulation of  
484 data (such as improper data augmentation) can lead to an overestimation of performance or lack  
485 of generalizability.

486

## 487 **3 Reference Standards**

488

489 The development of machine learning-based decision support tools requires truth or labeling  
490 of the cases for training, validation, and independent testing. The resulting reference standard  
491 needed for the evaluation of an algorithm’s (or human’s) performance depends on the task at  
492 hand. It is important to note that the notion of “truth” (or “ground truth” or “gold standard”) has  
493 been replaced by the concept of “**reference standard**,” as very few, if any, real-world tests yield  
494 the absoluteness implied by “truth” or “gold standard.” In many respects, the clinical utility of an  
495 algorithm greatly depends on the quality of the reference standard used in its training and  
496 evaluation. It is challenging but crucial for investigators to (1) select the most appropriate  
497 approach to obtain a task-specific reference standard, (2) gather complete and reliable data for  
498 that reference standard, and (3) assess any biases that may be introduced when training their  
499 algorithm with a reference standard that contains inherent variability.

500 This section covers considerations for generation of reference standards including objective  
501 vs. subjective reference standards, annotation granularity, methods for acquiring annotations,  
502 definition of true positives. The use of the reference standard in training and model development  
503 (section 4) and in performance evaluation (section 5) of a CAD-AI algorithm are closely related.

504

### 505 **3.1 Objective vs. Subjective Reference Standards**

506 The most straightforward reference standard uses the collected image data itself, with one or  
507 more domain experts providing diagnostic assessments or annotations at the image or patient

508 level. **Reference standards based on physicians’ opinion, however, are subjective**, and  
509 several studies have shown that CAD-AI system performance may vary substantially when  
510 assessed against different reference standards provided by radiologists [52-57]. Subjective  
511 reference standards are considered more reliable if they are based on consensus of multiple  
512 experts; however, it is difficult to estimate the number of experts needed. Ideally more than two  
513 experts should participate in order to identify outliers. It can be expected that the preferred  
514 number of experts depends on the task for which the reference standards will be used, the  
515 difficulty of that task, and the expected variability of the generated reference standard. In  
516 practice, obtaining a reference standard from experts is a very resource-intensive task, and  
517 usually only limited expert readings are possible, especially for large datasets.

518 Further reliability for reference standards may be achieved with information from other  
519 independent sources [58, 59], which also may be consensus based, such as radiologist’s review of  
520 images from another modality [60] or imaging follow-up for 2 years or longer [61].

521 Despite the prevalence of subjective approaches that use expert opinion, more **objective**  
522 **reference standards** are frequently desirable. For example, for lesion detection and pathologic  
523 classification, more definitive diagnostic tests and pathologic assessment of biopsied or excised  
524 lesions [62], although imperfect, should be used. For clinical decision support, such as treatment  
525 response assessment or patient prognosis, a more objective reference standard is patient survival.  
526 While the date of patient death is definitive, procuring this information as a reference standard  
527 becomes complicated by the need to track patients over potentially extended periods of time,  
528 during which they might become lost to follow-up; patient death could also result from  
529 circumstances other than the disease being evaluated. Shorter-term reference standards such as  
530 time-to-progression also may be used as an alternative in many studies.

531

### 532 **3.2 Annotation Granularity**

533 The level of required **annotation granularity**, or detail, depends on the task. For example, a  
534 more object-specific annotation such as manual expert delineation may be needed for  
535 lesion/organ detection or segmentation. For diagnosis of systemic disease or patient prognosis,  
536 patient-level assessment or patient survival may be appropriate. Image-based reference standards  
537 of varying levels of granularity are the most commonly used ones for current medical imaging-  
538 based machine learning tasks.

539

#### 540 **3.2.1 Entire image**

541 The coarsest level of granularity is **annotation of the entire image**, through which a class  
542 label is assigned to each image. As an example, the DREAM Challenge [63] for digital  
543 mammography diagnosis only had available breast-level labeling regarding the presence of breast  
544 cancer; however, training with such global labels that do not locate the actual lesions is sub-  
545 optimal in guiding deep networks to learn the relevant features of those lesions that are  
546 responsible for the patient-level diagnosis<sup>1</sup>. The top-scoring teams in the DREAM Challenge all  
547 used additional datasets with lesion location labeling to supplement the training of their systems.  
548 Another study showed that without specific lesion locations, the system could learn non-medical  
549 features that were included in the images (such as metal labels and markers), thus impeding the  
550 generalizability of the algorithm [11]. A more recent study [64] showed that the performance of  
551 an AI system for screening mammography on unseen cases varied from modest to outstanding  
552 depending on the dataset and reference standards used for evaluation.

---

<sup>1</sup> Recent “weak learning” and “attention” mechanisms may provide solutions for this (see Section 4.2)

553  
554  
555  
556 **3.2.2 Region-based**  
557 A finer level of granularity is annotation of specific lesions or organs through expert manual  
558 marking of a bounding box or a region center point. If the purpose is to detect cancers, for  
559 example, the CAD-AI system has to characterize the level of suspicion of a potential target  
560 structure and mark it as a cancer if it satisfies a certain threshold suspicion level. The scoring of  
561 system performance, then, requires not only the location of the lesion as reference standard but  
562 also the established malignancy status.

563  
564 **3.2.3 Pixel-based**  
565 An even finer level of granularity is **pixel-based annotation** in which the reference standard  
566 is an expert manual delineation, or outline, of the lesion or organ of interest and each image pixel  
567 can be labeled as either belonging to the region of interest or not. These detailed annotations are  
568 important for evaluating performance when the task is organ or lesion segmentation, and they can  
569 also be important for applications such as lesion characterization or treatment response  
570 assessment, in which the lesion extent and radiomic features are extracted from the segmented  
571 lesion. Pixel-based reference standards are more detailed than region-based ones but come at the  
572 cost of a more time-consuming annotation process and larger inter-reader variability [65].

### 573 574 **3.3 Methods for Acquiring Annotations**

575  
576 **3.3.1 Expert labels**  
577 When clinical or pathologic information is not available, it is common (for certain CAD-AI  
578 tasks such as lesion detection or segmentation) to create a **subjective reference standard from**  
579 **human domain experts**, who label the images or mark individual pixels, depending on the level  
580 of annotation granularity required. Outlining the boundaries of lesions or organs has the  
581 disadvantage of requiring potentially extensive time and effort, especially for manual  
582 segmentations in 3D. The judgment of lesion boundaries or the presence of a lesion contains  
583 intra- and inter-observer variability, even for experienced radiologists [65, 66], so that multiple  
584 experts may be required to produce a reliable reference standard.

585  
586 **3.3.2 Electronic health record**  
587 For patient-level assessments, the **electronic health records (EHR)** of subjects can be parsed  
588 by humans or natural language processing algorithms for reference standards involving, for  
589 example, the presence or absence of disease. Reference standards obtained from EHR data may  
590 contain annotations made during clinical practice, such as bounding boxes or Response  
591 Evaluation Criteria in Solid Tumors (RECIST) measurements [67]. If performed manually, a  
592 reference standard obtained from the EHR is time consuming and may not be practical for  
593 collecting large datasets; if performed automatically, the labels may contain a lot of noise and be  
594 prone to error, especially for complicated cases [68]. Natural language processing for parsing  
595 EHR data is an area of active research. It should be noted, however, that clinical radiology  
596 reports are not recommended as a reference standard for CAD-AI development, because “clinical  
597 reports often have nuanced conclusions and are generated for patient care and *not* for research  
598 purposes” [69].

### 599 600 **3.3.3 Crowd sourcing**

601 The key concept of **crowd sourcing** is to switch the time commitment and required effort for  
602 a given task from domain experts to many, potentially less-experienced, users. Crowd sourcing is  
603 a form of subjective consensus reference standard that has been applied to image annotation,  
604 image segmentation, and object delineation tasks [70]. It has been shown, in certain settings, that  
605 the quality of annotations from experts and those from novices becomes equivalent with an  
606 increased number of novices [71, 72]. Nevertheless, the use of crowd sourcing as a reference  
607 standard for machine-learning applications in medical imaging must be further investigated  
608 before it can be recommended for general use.

609

### 610 3.3.4 Phantoms

611 In medical imaging, **phantoms** are man-made objects with known structure and composition.  
612 Images acquired of these phantoms support *a priori* image annotations across a range of  
613 granularity levels. However, the number of physical phantoms usually is limited, and, therefore,  
614 only a few annotated images can be obtained from this method. Recently, digital phantoms that  
615 mimic properties of physical objects *in silico* have become available [73] and have been used in  
616 virtual clinical trials [73, 74] as well as for training ML models [39]. An advantage of using *in*  
617 *silico* models is that the lesion location and properties are known by design so that human  
618 annotation is not required; however, image data obtained from phantoms (physical or digital)  
619 likely do not reflect the actual biological or pathological characteristics that may be captured on  
620 patient images. Phantom images may be useful for data augmentation during training, for  
621 identifying and correcting biases regarding differences in imaging systems and protocols, and for  
622 test-retest evaluations. Whether an algorithm trained with phantoms is applicable to real-world  
623 images requires rigorous validation [39]. Similar caution must be applied to the use of synthetic  
624 images generated by digital methods such as full *in silico* modeling of the imaging chain or use of  
625 generative adversarial networks.

626

### 627 3.3.5 Weak/noisy labels

628 **Weak or noisy labels** can be defined as incomplete or imperfect reference standard  
629 annotations. Compared with a small dataset with “strong” or “clean” labels, a large dataset with  
630 “weak” or “noisy” labels used for algorithm training may achieve comparable performance [72].  
631 The generalizability of the trained algorithm, however, will deteriorate as the proportion of noisy  
632 labels in the training set increases [75]. Others have demonstrated the potential of using weak or  
633 noisy labels [76] but additional research is needed. Strong labels specifically for the independent  
634 test set are essential to reliably assess the performance of the trained decision support tool. Under  
635 certain circumstances, the STAPLE algorithm (“Simultaneous Determination of a Reference  
636 Standard and Performance Level Estimation”) delivers not only the optimal reference standard  
637 estimation but also a quality ranking of the competing observers/algorithms [77].

638

### 639 3.4 Definition of True Positives

640 Reference standards are designed for use in evaluating the output of a CAD-AI system. The  
641 definition of a **true positive** relative to the reference standard is very important. Different  
642 methods for determining a true positive will result in different performance of the same CAD-AI  
643 system. Which method is appropriate or feasible depends on the task and the available reference  
644 standard. Using detection tasks as a specific example, a number of methods have been used to  
645 determine whether the lesion is correctly detected, including the distance between the centroids  
646 of the detected object and the reference, the overlap percentage between the two (which is further  
647 affected by the level of detail in marking the reference, e.g. bounding box vs. outline) [78], and  
648 whether the centroid of the detected object falls within the reference lesion region; detected

649 objects that are not determined to be true positives through the selected metric are counted as  
650 false positives. It has been shown that scoring is strongly affected by the detection criterion [79].  
651 More detail on performance evaluation can be found in section 5.

652  
653 **3.5 Take Home Message on Reference Standards**  
654 The required type and granularity of the reference standard depends on the task at hand. An  
655 objective reference standard is preferred; however, when a subjective reference standard cannot  
656 be avoided, independent assessments of multiple domain experts should be obtained and their  
657 variabilities should be evaluated.

658

## 659 **4 Model Development**

660  
661 In addition to the availability of properly collected data and labels, the selection of data  
662 sampling and machine learning strategies will affect the robustness of the developed model. This  
663 section covers the topics of data sampling methods, levels of learning supervision, and new  
664 training strategies, including transfer learning, multi-task learning, domain adaptation, federated  
665 learning, and continuous-learning. A recent review on some of these technologies and their  
666 applications can be found in the literature [80].

667

### 668 **4.1 Data Sampling Strategies**

669 Data sampling is important for efficient use of data and for reducing the risk of overfitting in  
670 model development. The most established resampling techniques for the training and testing of  
671 models will be discussed. The dataset ideally should be split into three non-overlapping  
672 partitions: **training**, **validation**, and **test** sets. One of the partitions should be used for training of  
673 the model. To guide the optimization (or tuning) of model parameters during training of a model,  
674 it is desirable to obtain a meaningful estimate of the performance of the model being trained on a  
675 partition of the dataset that is often referred to as a “validation set;” the use of the validation set is  
676 thus a part of the training process. This is not to be confused with the use of the term  
677 “validation” as the process of evaluating the generalizability of a developed model on unseen  
678 data after training is completed and the model is “frozen,” which should be established by **testing**  
679 **on a completely independent dataset** from the ones used during the training or optimization of  
680 the model. To avoid overfitting the model, performance testing ideally should be conducted only  
681 once on any given **test set**; the performance on that test set should then not be used to inform  
682 model improvements or modifications for subsequent testing on the same test set [5, 14, 81]. Due  
683 to potential confusion surrounding the term “validation” for reporting the performance of a  
684 trained model, developers need to clearly define whether the test set used for the evaluation has  
685 been kept independent from the training process. There are several established resampling  
686 techniques for organizing the training and evaluation of a model, especially with limited datasets.  
687 It should be noted that such techniques are generally based on the assumption that the available  
688 data are representative of the underlying target population and similarly distributed within the  
689 training, validation, and test datasets.

690 A **holdout method** is the most basic evaluation/training paradigm. In this approach, a model  
691 is trained and optimized by use of training and validation datasets, after which it is evaluated  
692 once with an independent test dataset that is sequestered during training. When the available  
693 datasets are small, a **k-fold cross-validation** method, which maximizes the use of the available  
694 data, can provide a more reliable evaluation of model performance than the holdout methods

695 under this condition [82, 83] if the test partition in each fold is held-out as an independent test set  
696 and is not used repeatedly for guiding model optimization. For such techniques, stratified  
697 sampling of cases (Section 2.1) can better accommodate imbalanced datasets than random  
698 sampling. **Bootstrapping** is another popular and well-established resampling method that can be  
699 used to construct sampling distributions for model training and evaluation purposes [84-86].

700 Although the actual generalization performance of the final model should be evaluated only  
701 once by external testing with a previously unseen independent test set, in practice, it is  
702 psychologically difficult for researchers not to go back and improve their model if the observed  
703 test performance is poor. Such multiple testing and reuse of the same test data are likely to  
704 introduce overfitting problems regardless of the evaluation/learning paradigm [81, 87].  
705

## 706 **4.2 Machine Learning Strategies**

707 A machine learning paradigm refers to a strategy based on which a model is trained. There  
708 are numerous learning paradigms in CAD-AI, many of which overlap [88-90]. One approach for  
709 categorizing learning paradigms focuses on the level of interaction required by the user, such as  
710 supervised, semi-supervised or unsupervised learning. A different approach considers the  
711 learning paradigm from the perspective of model development, such as transfer learning, multi-  
712 task learning and federated learning.  
713

### 714 **4.2.1 Levels of learning supervision**

715 Supervised learning (with different levels of supervision) is the most common approach to  
716 learning, where a model is trained to map input data to output data based on examples of the  
717 input-output pairs. To reduce the cost and barriers related to data collection and annotation,  
718 however, several studies are actively exploring training algorithms that can leverage unlabeled or  
719 weakly labeled data during training (see also Section 3.3.5). Such paradigms may provide a more  
720 cost-effective and scalable approach to CAD-AI development.  
721

#### 722 **4.2.1.1 Supervised learning**

723 In **supervised learning**, a model is trained to map input data to output data based on explicit  
724 examples of the desired input-output pairs, as provided by the user. However, the collection of  
725 such annotations tends to be costly and time-consuming, and the annotation effort may need to be  
726 repeated as the imaging technology evolves and new datasets are introduced. Moreover, as noted  
727 in previous sections, annotations can be subjective, the annotation process may be prone to error,  
728 and, for some tasks, there is no single correct annotation.  
729

#### 730 **4.2.1.2 Semi-supervised learning**

731 **Semi-supervised learning** algorithms exploit a combination of labeled and unlabeled data. In  
732 this case, the model is given some guidance about the desired outcome, but the annotations do not  
733 need to be as detailed or extensive as those used with supervised learning. For instance, feature  
734 extraction can be initialized through an unsupervised or self-supervised technique and then fine-  
735 tuned to the final task using a small set of labeled data. Using some form of semi-supervised  
736 learning may reduce the costs of labeling relative to supervised learning.  
737

#### 738 **4.2.1.3 Self-supervised learning**

739 **Self-supervised learning** can exploit large unlabeled datasets for feature representation and has a  
740 regularizing effect on the learned features. Autoencoder models are a common approach to self-  
741 supervised learning [37] and are used for feature extraction; however, there is no guarantee that

742 the features learned in a self-supervised fashion have diagnostic value. It should be noted that  
743 autoencoder models, such as U-Net, can also be used in a supervised mode for image  
744 segmentation tasks. Other popular approaches to self-supervised learning include *contrastive*  
745 *learning* [91-93] and *pretext* [91] or *surrogate supervision* [94]. In these techniques, when a large  
746 unlabeled dataset in the same domain as a small labeled dataset is available for a given task, the  
747 unlabeled data can be assigned artificial labels and then used to pre-train a deep learning model;  
748 transfer learning for the target task is then performed with the small labeled dataset. It has been  
749 shown that deep models pre-trained with self-supervised learning techniques can outperform the  
750 same models trained with random initialization [95] or transfer learning from an unrelated  
751 domain [94, 96]. These findings demonstrate the potential of large datasets to improve model  
752 development in medical imaging tasks even if a large portion of the cases is unlabeled.

753

#### 754 4.2.1.4 *Unsupervised learning*

755 **Unsupervised learning** refers to a class of algorithms that can autonomously learn from data  
756 without reference to any labels or any instruction from the user. Common approaches to  
757 unsupervised learning are the clustering methods. Unsupervised learning has shown promise in  
758 medical imaging applications but depends on the adequacy of the resulting automatic clustering.  
759 In addition, unsupervised learning requires a much larger training set for the algorithm to achieve  
760 similar performance compared with training with reference standard [97], and data collection in  
761 medical imaging is costly.

762 It should be noted that CAD-AI algorithms can include both supervised and unsupervised  
763 elements.

764

#### 765 4.2.1.5 *Multiple-instance learning*

766 The **multiple-instance learning** approach is an effective paradigm when labels are not  
767 available at the desired granularity [98]. The machine learning model receives a set of labeled  
768 “bags,” each containing many (unlabeled and some labeled) instances. In the simplest case of  
769 binary classification, a bag is labeled positive if it contains at least one positive instance.

770

### 771 4.2.2 **Transfer learning, multi-task learning, and domain adaptation**

772 The ability to discover by **representation learning** a wide range of object characteristics is a  
773 distinctive advantage of deep learning over traditional machine learning models that rely on  
774 hand-engineered features [99]. In deep convolutional neural networks (DCNNs), feature  
775 extraction is obtained through a series of cascaded convolutional layers, forming a hierarchy in  
776 which shallow layers extract generic features and deeper layers extract increasingly object-  
777 specific features [100]. Large-scale datasets, however, are needed to learn high-quality features,  
778 thus making deep learning an effective, but data and computation hungry, paradigm. Such data  
779 requirements can be lessened by transferring or sharing features across different tasks and  
780 domains.

781

#### 782 4.2.2.1 *Transfer learning*

783 **Transfer learning** in DCNNs is commonly implemented by training a network on one task  
784 and then “transferring” the parameters (or weights) from the trained model to initialize the  
785 network for a new task, rather than randomly initializing it (also known as “training from  
786 scratch”). Transfer learning was the early enabler for the use of deep networks in the medical  
787 imaging domain. Networks pre-trained on ImageNet, which comprises millions of non-medical  
788 images effectively labeled by crowd sourcing, are commonly used to initialize DCNNs for

789 medical image classification, showing improved classification performance and faster  
 790 convergence compared with random initialization [98, 101-105]. Transfer learning, however,  
 791 imposes limitations on the DCNN, since ImageNet is composed of low-resolution 2D RGB color  
 792 images, whereas many medical imaging modalities are higher-resolution 3D, 4D, or multi-  
 793 parametric. One of the most common techniques for bridging the two domains involves a 2.5D  
 794 approach [106], in which a 3D (or higher-dimensional) image around a voxel is subsampled into  
 795 multiple 2D images, which are then fed into the input channels of a 2D DCNN [102] or an  
 796 ensemble of 2D DCNNs [107].

797 For some tasks, such as segmentation, 3D convolutional filters may offer substantial  
 798 advantages over 2D CNNs; in such cases, training from scratch or transfer learning from another  
 799 medical imaging modality may be performed. Researchers have begun to explore medical  
 800 imaging-based pre-training of DCNNs, and results indicate that an additional stage of pre-training  
 801 with data from a similar domain can increase performance and robustness of a network [108,  
 802 109]. The transfer of prior knowledge can occur between modalities (e.g., CT to MRI), between  
 803 organs/pathologies (e.g., liver to kidney), between tasks (e.g. classification to segmentation), or  
 804 some combination thereof [110].

805

#### 806 4.2.2.2 Multi-task learning

807 **Multi-task learning** is a special type of transfer learning in which a DCNN is trained to  
 808 jointly learn interrelated tasks, as opposed to addressing each task sequentially [111]. This  
 809 technique has demonstrated enhanced performance compared with single-task learning [110,  
 810 112].

811

#### 812 4.2.2.3 Domain adaptation

813 Most algorithm training methods assume that the test data is drawn from the same distribution  
 814 as the training data; however, this assumption is often not fulfilled in practice due to data scarcity  
 815 and data mismatch, and thus a trained model may fail to generalize to real-world clinical data  
 816 [113, 114]. The most important sources of **data shift** (i.e., deviations between the distributions of  
 817 the test set data and the training set data) in medical imaging are acquisition shift and population  
 818 shift (Table 1) [11].

819 Data shift can be addressed, at least partially, through data harmonization and standardization,  
 820 as discussed in Section 2.3. Recently, researchers in the medical imaging space have begun to  
 821 explore domain adaptation techniques to make deep learning models more tolerant of domain  
 822 shift [115]. The most common approaches to domain adaptation are feature based and attempt to  
 823 modify the feature distributions to align the target (i.e., test set) and source (i.e., training set)  
 824 domains. Other approaches seek to learn domain-invariant representations [116] or use generative  
 825 models to synthesize realistic samples in target domains where labeled data are scarce [117-120]  
 826 [38].

827

828 Table 1. Type of data shift.

<b>Data Shift</b>	<b>Definition</b>
<b>Prevalence shift</b>	training and test datasets have different disease prevalence (class imbalance)
<b>Acquisition or domain shift</b>	different imaging equipment or imaging protocols are used between training and test datasets
<b>Population shift</b>	intrinsic characteristics (e.g., demographics or disease presentation) of the populations under study differ between training and test datasets

<b>Annotation or label shift</b>	class definition changes between training and test datasets, e.g., due to inter-rater variability or lack of standardization in the class definitions
----------------------------------	---

829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873

### 4.2.3 Federated learning

**Federated learning** is a distributed machine learning approach that enables collaborative training on decentralized datasets [121-124]. Each site trains the model locally with its own dataset and then only the trained model parameters are shared, thus producing a global model benefiting from access to a large corpus of data without requiring data sharing and without posing risks to patient privacy. There are, however, several open-ended questions with regard to federated learning that are relevant to medical imaging [125, 126]. In particular, there is no formalized training protocol yet to guarantee that the performance of a model trained with federated learning is comparable to that of a centralized trained model with access to all the data [127]. Also unknown is (1) the extent to which local model overfitting negatively impacts the global model, and (2) the tradeoff between access to more data through a federated process versus traditional learning with a fully controlled dataset.

### 4.2.4 “Continuous learning” systems

**Continuous or “life-long” learning** emulates the human ability to continuously learn and adapt as new data are presented [128, 129]. Theoretically, continuously learning AI systems can accelerate model optimization and continuously improve their performance by taking advantage of new data presented during clinical use. In practice, adaptive training of shallow and deep neural networks using incrementally available data generally results in rapid overriding of their weights, a phenomenon known as “interference” or “catastrophic forgetting” [130, 131]. It is not generally clear under what conditions and for what metrics adaptive AI produces a continuously improving (or at least stable) algorithm and avoids major pitfalls. Many questions related to post-marketing management of adaptive AI devices remain open, such as frequency of adaptation (e.g., continuously or in regular intervals, batch mode), how to monitor performance changes after adaptation, and when and how to intervene if performance decline is suspected.

## 4.3 Take Home Message on Model Development

Training approaches, especially for deep learning algorithms, are continuously improving with the goal of achieving robust, effective, and privacy-preserving CAD-AI models. An independent test set representative of the intended use that was not employed to guide model optimization in any learning paradigm is of critical importance. Robust training methods, although important for all CAD-AI systems, are especially important for systems that may operate in clinical practice with minimal or no human supervision.

## 5 Performance Assessment

Proper performance assessment is important in various stages of CAD-AI model development. Performance assessment involves (1) factors such as intended use, performance metrics, statistical significance, sample size, and reproducibility and (2) purposes such as standalone or clinical reader performance assessment. Rigorous performance assessment can provide a reliable estimate of model performance at a particular stage of development to guide further improvement or to inform the user of realistic performance that can be expected from the model. This section discusses methods and considerations for conducting performance assessments.

874

## 875 **5.1 Performance Assessment Metrics**

876 In CAD-AI applications, the most widely accepted performance assessment methodologies  
877 include receiver operating characteristic (ROC) analysis [132], its various derivatives such as  
878 free-response ROC (FROC) analysis [133], and precision-recall analysis. In detection and  
879 classification tasks, the most common metrics include area under the ROC curve, sensitivity (or  
880 recall), specificity, balanced accuracy (mean of the sensitivity and specificity), Youden index,  
881 and the prevalence-dependent factors positive predictive value (or precision), negative predictive  
882 value, and F1 score [5, 134, 135]. Various other methodologies and metrics have been established  
883 for specific applications, such as the Dice coefficient, Jaccard index, and Hausdorff distance for  
884 image segmentation; mean squared error and coefficient of determination for regression;  
885 concordance index [136, 137] for evaluating prediction performance; the log-rank test [138] for  
886 comparing Kaplan-Meier survival curves in survival analysis; and categorical agreement of  
887 response classification by, for example, the RECIST guidelines [139, 140]. The use of multiple  
888 performance approaches is generally appropriate to provide a more complete assessment.

889 It is crucial to include error estimates, such as standard deviations or 95% confidence  
890 intervals, when reporting results. Error estimates describe the uncertainty/variability of the  
891 reported values for the performance metrics and help provide insight into the sufficiency of the  
892 training sample size, the soundness of the training/testing approach, and generalizability.

893

## 894 **5.2 Statistical Significance**

895 **Statistical significance** is used to quantify the likelihood that an observed result is  
896 explainable due to chance alone [141]. **Statistical power** is a closely related topic that quantifies  
897 how likely a study is to distinguish an actual effect from one of chance. Whereas statistical  
898 significance of results is assessed *after* study completion, statistical power calculations are an  
899 important part of study design and performed *beforehand* to estimate the required sample size  
900 based on the expected size of the effect, variability in the response variable, and disease  
901 prevalence [142]. Failure to achieve a statistically significant result cannot be interpreted as a true  
902 lack of difference especially when the study is statistically underpowered. It is important to note  
903 that statistical significance does not necessarily imply that the result is clinically meaningful  
904 [143, 144] unless the study is specifically powered to address this issue. Moreover, when multiple  
905 statistical hypotheses are tested using the same dataset, the chance of observing a rare event  
906 increases, thereby increasing the likelihood of incorrectly concluding that a real effect has been  
907 observed when the observation, in fact, was due to chance alone; methods for adjusting for the  
908 effect of multiple hypothesis testing have been developed [145]. Statistical tests generally make a  
909 set of assumptions about the distribution of the data to which they are applied (e.g., normality or  
910 linearity), and it is important to verify these assumptions are met before using any specific  
911 statistical test.

912

## 913 **5.3 Intended Use**

914 The **intended use** for which a CAD-AI system is designed must match the clinical  
915 environment in which it is deployed. The intended use is determined by the patient population,  
916 the image acquisition device, the stage of diagnostic intervention, and the diagnostic category.  
917 First, the patient population represented by the data used to develop the algorithm should match  
918 the intended population. Second, a range of image acquisition devices are in clinical use, and  
919 CAD-AI must be developed and tested on data from multiple vendors. Third, the intended use  
920 depends on the patient care stage that requires the diagnostic intervention. Finally, the diagnostic

921 category of the data should match the clinical task, for example, screening, detection, staging,  
922 treatment assessment, or follow-up.

923 CAD-AI systems for aiding in clinical decision making generally may be implemented  
924 according to four different paradigms: **second read, concurrent read, triage, and rule-out.**  
925 CAD-AI applications such as detection and diagnosis as well as staging, treatment response  
926 assessment, prognosis, or recurrence prediction (Figure 1) should be matched with the most  
927 appropriate paradigm. The selected performance assessment method should be reflective of the  
928 use paradigm (Table 2). Frequently, the setting may affect the operating point of the CAD-AI  
929 tool, e.g., the relative importance of sensitivity vs. specificity. In addition, CAD-AI systems  
930 designed for different intended uses may have different performance requirements; for example,  
931 CAD-AI systems designed for disease detection in a concurrent-read paradigm generally should  
932 have higher sensitivity and specificity than those used in a second-read paradigm due to  
933 potentially increased reader reliance on the computer output in the former setting. CAD-AI  
934 devices that operate at performance levels that rival those of human experts [146-148] could  
935 potentially be the basis for future autonomous AI devices that bypass human interpretation in  
936 selected cases or for selected tasks. An example of such applications is **rule-out devices**, a class  
937 of devices designed to identify and remove negative cases without clinician review. Although  
938 some authors have considered rule-out as a subset of the triage paradigm, the clinical  
939 implementation of each requires a unique set of strategies and performance assessment  
940 considerations due to different levels of risk associated with each approach.

941

#### 942 **5.4 Standalone Performance Assessment**

943 The evaluation of a CAD-AI algorithm includes both benchmarking algorithm performance  
944 and assessing the added value to the end user provided by the algorithm in improving clinical  
945 decision making [5]. **Standalone performance assessments** are employed during development  
946 to allow for modifications to be quickly compared to previous models. For benchmarking,  
947 overall performance is based on an independent dataset representative of the clinical population  
948 acquired using the expected range of image acquisition technologies and protocols for the  
949 intended use.

950

951 Table 2. Different paradigms of CAD-AI systems.

<b>Paradigm</b>	<b>Intended Use</b>	<b>Evaluation approach</b>
<b>Second read</b>	Improving decision making by providing a second opinion to the physician <i>after</i> initial interpretation	Assessment of physician performance without and with the aid in a sequential reader study design; first interpret each case without, then with, CAD-AI system [5, 134, 149-151]; or independent or crossover study design similar to that of concurrent read.
<b>Concurrent read</b>	Improving decision making by showing system output to the physician <i>at the same time</i> as initial interpretation	Assessment of physician performance without and with the aid in an independent or crossover reader study design; cases are interpreted in batches either with or without the aid after a sufficient washout time and in counter-balanced manner to reduce the potential memorization effects [5, 134, 152]

<b>Triage</b>	Improving workflow by prioritization: All cases are interpreted but order prioritized by CAD-AI system [153, 154]	Assessment of process improvement by local clinical operations
<b>Rule-out</b>	Improving workflow by removal of normal or negative cases from workflow: The removed cases are not interpreted by physician.	Comparison of performance with and without rule-out in clinical practice [146-148]

952  
953  
954

### 955 **5.5 Clinical Reader Performance Assessment**

956 A **clinical reader performance assessment** is used to estimate the clinical impact of a CAD-  
957 AI algorithm [153, 154]. A common approach for assessing clinical performance is through a  
958 controlled reader study (either retrospective or prospective), directly comparing the performance  
959 of a human reader without and with output from the CAD-AI system [155, 156]. A disadvantage  
960 of this approach is that the estimated performances are unlikely to match those in the true clinical  
961 setting because of differences in the cases, physicians, and reading process. It is important to  
962 realize that both the population of patients undergoing the examination (cases) and the population  
963 of physicians interpreting the data (readers) are sources of substantial variability in clinical reader  
964 studies [157]. Specialized statistical and methodological tools are needed for these analyses  
965 [158]. Well-designed clinical reader studies can be used to gain Food and Drug Administration  
966 approval (or approval by a similar organization outside of the United States) for clinical use of a  
967 CAD-AI system and are often a precursor to direct assessment of diagnostic performance in  
968 clinical practice (Section 6.4.3).

969  
970

### 970 **5.6 Sample Size**

971 Assessing performance dependency on the training **sample size** in medical imaging is  
972 important to achieve a viable clinical translation. As previously discussed (Section 4.1), small  
973 training sample sizes may lead to overfitting, or overtraining, of CAD-AI algorithms. In general,  
974 the performance of CAD-AI systems depends on the training sample size, disease prevalence, the  
975 number of features and their statistical distribution, the choice of the CAD-AI model, and the  
976 scoring metric [82, 85, 159, 160]. For the deep learning techniques, the training sample size is  
977 even more critically important since millions of parameters need to be determined. Even when  
978 deep learning models are trained with transfer learning (Section 4), the training sample size is  
979 still a major factor that affects performance and generalizability. The variability in the algorithm  
980 performance from repeated experiments at different sample sizes can be used to assess overfitting  
981 and generalization error [75, 108].

982  
983

### 983 **5.7 Reproducibility**

984 It is important to clearly specify the conditions under which the results of a CAD-AI system  
985 are reproducible. Recent studies have distinguished among different types of reproducibility  
986 [161-163]. Three types of reproducibility have been defined, the first two of which are relevant  
987 for model validation and successful clinical deployment of CAD-AI systems.

988 **Technical reproducibility** refers to the ability to precisely replicate reported results (usually  
989 in a publication) based on a complete description of the method and release of the corresponding  
990 code and dataset.

991 **Statistical reproducibility** refers to a result being valid (within a specified standard deviation  
992 or confidence interval) when different variations of the training conditions are applied. Variations  
993 in training conditions will result, for example, from different random seeds, from different  
994 partitions of the training set, or from different strategies to divide the dataset into training and test  
995 subsets. Statistical reproducibility in model performance will also depend on the test set. If  
996 different test sets are sampled from the same population, the DCNN output will be different for  
997 the different test sets due to statistical variation of the test sets. If the test is repeated multiple  
998 times, and each time a different test set is randomly drawn from the population or by  
999 bootstrapping, the test performances can be considered samples from the same statistical  
1000 distribution, from which the mean performance and standard deviation can be estimated.

1001 **Inferential reproducibility** refers to the ability to reach qualitatively similar conclusions  
1002 from an independent replication of a study under conditions that match the conceptual description  
1003 of the original study.  
1004

## 1005 **5.8 Take Home Message on Performance Evaluation**

1006 The most appropriate performance metric(s) will depend on the task and the reference  
1007 standard. Often multiple performance metrics are appropriate, and use of multiple metrics is  
1008 frequently desirable. Power calculations should be an integral part of study design, and  
1009 performance analysis should include error estimates, assessment of statistical significance, and  
1010 preferably assessment of reproducibility.  
1011  
1012

## 1013 **6 Translation to Clinic**

1014  
1015 The ultimate goal of developing CAD-AI system is to assist physicians in the health care process.  
1016 For clinical acceptance of a CAD-AI tool, many practical factors must be considered, such as  
1017 generalizability to the clinical environment, efficiency of use in a clinical workflow,  
1018 explainability of the output, and assurance of performance consistency over time. This section  
1019 will discuss topics related to the translation of CAD-AI tools to the clinic, including human-  
1020 machine interface, user training, acceptance testing, and prospective surveillance.  
1021

### 1022 **6.1 Human-Machine Interface**

1023 One of the most important issues of introducing CAD-AI to clinical use is the presentation of  
1024 its output to the physician. The **human-machine interface** is a critical component that can  
1025 impact the usefulness and the acceptability of a CAD-AI tool for clinical use. The interface  
1026 design will depend on the intended use (e.g., disease detection, triaging, treatment response  
1027 assessment); the amount, type, and complexity of information to be displayed (e.g. markers,  
1028 parametric maps, likelihood scores); the reader paradigm; and the level of interactivity (e.g.,  
1029 when and how the physician can enable, disable, or query the CAD output). Regardless of the  
1030 task, some common requirements may include user friendliness, workflow efficiency, and the  
1031 interpretability of the CAD-AI output or recommendations.

1032 The black-box nature of current CAD-AI tools is one of the roadblocks to translation of  
1033 CAD-AI into clinical use. Providing uncertainty estimates of the output could allow a better  
1034 understanding of the black box model and improve the safety of deep learning systems [164-168].  
1035 For physicians to have confidence in a recommendation by a CAD-AI tool, it is helpful for them  
1036 to understand the reasons behind the prediction or decision. The explanation has to be consistent  
1037 with medical knowledge or supported by clinical evidence. **Explainable AI (XAI)** is an

1038 emerging machine learning area [169] that seeks to design interpretable AI models or, more  
1039 commonly, provide post-hoc explanations for trained AI models; the most common approaches at  
1040 present include generating visual heatmaps, providing examples of similar lesions or cases, and  
1041 providing semantic textual explanations or cues [170]. A visual saliency map or a color heatmap  
1042 of the image [171], which captures the relative contribution to the DCNN output score from  
1043 various image locations, can be generated using a gradient-based, perturbation-based, or class  
1044 activation map-based (CAM) method [172-176]. The local interpretable model-agnostic  
1045 explanations method (LIME) [177] similarly identifies the extent to which regions or pixels  
1046 influence the particular prediction. The visualization provides some evidence of the correlation of  
1047 the deep features and the output score to the input data; however, visualization maps (which are  
1048 generally difficult for humans to interpret) are far from a complete explanation of why and how  
1049 the features are connected and weighted to identify the target lesion [169, 176]. Saliency map  
1050 techniques often cannot meet key requirements for utility and robustness, emphasizing the need  
1051 for additional validation before clinical use [176]. For clinical tasks more complicated than lesion  
1052 detection, the CAD-AI tool may need to provide explanations or references that correlate the  
1053 recommendation with the patient's medical conditions or other clinical data. Much more research  
1054 and development are needed to determine physicians' preferences regarding user interface design  
1055 for each type of application so that CAD-AI models can truly become intelligent decision support  
1056 tools.

1057

## 1058 **6.2 User Training**

1059 In translating technology to the clinic, an important step is to set expectations. Key to a **user's**  
1060 **proper use of a CAD-AI tool** is an understanding of the intended use, including the purpose and  
1061 when and how it should be used in the radiology workflow [178]. For example, if a CAD-AI tool  
1062 is developed for lesion detection, the user should be informed about whether the tool is designed  
1063 and validated for use in a concurrent-read or second-read paradigm. CAD-AI tools designed for  
1064 different intended uses may have different performance requirements; for example, CAD-AI  
1065 systems designed for disease detection in a concurrent-read paradigm generally should have  
1066 higher sensitivity and specificity than those used in a second-read paradigm due to potentially  
1067 increased reader reliance on the computer output in the former setting.

1068 A second key issue is to **acquaint the user with both the capabilities and limitations of a**  
1069 **specific decision-support tool**. Users should have a comfortable level of trust in the CAD-AI  
1070 tool but should always be aware of the performance limitations of the tool. The performance of a  
1071 CAD-AI tool can be affected by patient demographics, imaging equipment, and image-  
1072 acquisition protocols. Even if a CAD-AI tool has been trained by the vendor with multi-  
1073 institutional data and approved for clinical use, its performance in the local population may not  
1074 be the same as that specified by the vendor. An initial user-training and adjustment phase is  
1075 recommended as an integral part of the deployment. During this phase, physicians should  
1076 evaluate the performance of the CAD-AI tool on their patient cases by comparing with clinical  
1077 outcomes to understand the characteristics of the cases for which the CAD-AI provides correct  
1078 and incorrect recommendations, but they should refrain from being influenced by the CAD-AI  
1079 output in their clinical decisions. This adjustment phase will provide the user with a deeper  
1080 understanding of the CAD-AI performance in the local setting, and also impart to the user an  
1081 appropriate level of confidence in the recommendations generated by the decision-support  
1082 system, which may reduce unrealistic expectations and improper use of a CAD-AI tool. For  
1083 example, misusing a tool intended to be a second opinion as a concurrent reader may lead to  
1084 disappointing outcomes, user dissatisfaction, and, most importantly, potential harm to patients  
1085 [179]. The length of this training period may depend on the type of CAD-AI application, the level

1086 of risk, and the observed performance and consistency of the CAD-AI tool. The resulting insights  
1087 may also provide useful feedback for the CAD-AI vendor [6].  
1088

### 1089 **6.3 Acceptance Testing**

1090 CAD-AI software to be implemented for clinical use is considered a medical device; its  
1091 performance, therefore, must meet certain standards. **Acceptance testing** is an important step  
1092 prior to clinical use of any CAD-AI tool [6, 178]. Manufacturers must provide instructions for  
1093 use with detailed guidance on system installation, acceptance testing, acceptance criteria at  
1094 installation and subsequent upgrades, and periodic QA. The instructions for use must also  
1095 include a description of the expected performance levels of the CAD-AI system along with  
1096 tolerance limits and a graphic presentation of CAD-AI output layout and proper user interface  
1097 configuration.

1098 A **basic level of acceptance testing** may use pre-collected data provided by the manufacturer  
1099 or phantoms for testing the operation and consistency of certain CAD-AI functions after  
1100 installation and compared with the expected outcomes. **Another level of acceptance testing**  
1101 should include a set of clinically representative cases collected by the individual clinical site.  
1102 The deviation of the resulting performance level from the performance level claimed by the  
1103 CAD-AI manufacturer must be within specified tolerance limits. For clinical sites that may not  
1104 have a large set of cases readily available for acceptance testing, the clinical performance  
1105 assessment may be conducted during the user training phase, which may be less quantitative but  
1106 has the advantage of being most consistent with the clinical operations at that site.  
1107

### 1108 **6.4 Prospective Surveillance**

#### 1109 **6.4.1 Periodic quality assurance**

1110 The goal of **periodic QA** is twofold: to establish a schedule of routine QA and to assure the  
1111 consistency of clinical performance over time. Routine QA should be implemented (preferably  
1112 by medical physicists in conjunction with routine QA testing of related medical imaging systems)  
1113 to assess how variations in the imaging or data collection chain may affect the performance of the  
1114 CAD-AI system [6, 178]. QA should also be performed whenever a CAD-AI software update  
1115 occurs, which should always be announced by the software development company. The use of  
1116 phantoms for this testing is recommended if the CAD-AI system is designed to be applicable to  
1117 specific phantoms and its performance has been shown to be sensitive to the quality of images  
1118 acquired from these phantoms. To evaluate performance consistency in routine clinical cases,  
1119 clinical sites and CAD-AI manufacturers should develop tools to track performance levels of  
1120 certain indices and monitor deviations (e.g., a tool to track the number of markers per image for  
1121 detection tasks [6]).  
1122

1123 The tolerance limits and corrective actions for any observed deviations should be established  
1124 based on the CAD-AI application. The risk associated with any deviation will vary significantly  
1125 for different diseases and tasks performed by the CAD-AI system. For example, if the system is  
1126 an autonomous CAD-AI detection or decision tool for triaging or rule-out, immediate corrective  
1127 actions are recommended, while tools designed only to provide second opinion or supplementary  
1128 information may be less urgent. Regardless of the risk level, awareness of these deviations by the  
1129 physicians is critical as they may need to adjust their level of trust on the CAD-AI  
1130 recommendation when performing clinical tasks.  
1131

#### 1132 **6.4.2 Performance monitoring for “continuous learning” systems**

1133 For continuous learning CAD-AI systems implemented in the clinic, an additional risk results  
1134 from learning from non-stationary data that may lead to catastrophic forgetting and degraded  
1135 performance unbeknownst to the physicians in their daily use of the system [129]; furthermore,  
1136 system performance may be frequently changing, which impacts its safety profile. The  
1137 manufacturer or the in-house development team must have well-defined QA procedures to  
1138 validate the quality of data, including collateral information (e.g., clinical outcomes), and assess  
1139 model performance after each update. Before continuous learning CAD-AI systems can be  
1140 translated into the clinic, extensive work is required to develop practical and reliable QA methods  
1141 that enable performance monitoring to ensure safe use.

1142

### 1143 **6.4.3 Prospective evaluation of CAD-AI**

1144 Large-scale prospective performance assessment of CAD-AI systems will evaluate the impact  
1145 of CAD-AI on workflow efficiency, physician performance, cost-effectiveness, and patient  
1146 outcomes in the clinical setting. Prospective evaluation of CAD-AI typically falls into two  
1147 categories: **randomized controlled trials (RCTs)** and **observational studies**.

1148 **RCTs** are designed to control for sources of bias through randomization, blinding, and  
1149 allocation concealment. RCTs are logistically difficult to organize and generally require a large  
1150 patient population. A common design is the sequential study, in which the physician interprets  
1151 each case first without the assistance of CAD-AI and then, after formally recording his or her  
1152 findings, interprets the case again while reviewing the CAD-AI recommendation [180-186]. This  
1153 sequential design, however, cannot be applied with concurrent-read or triage paradigms, as  
1154 discussed in Section 5.3 (Table 2).

1155 Well-designed **observational studies** can be highly informative and much easier to conduct  
1156 than RCTs [187]. The most common design is the historical-control study, in which the  
1157 performance of groups of radiologists over different periods of time is compared; the patient  
1158 cohorts and radiologists involved may not be identical for the two time periods. Observational  
1159 studies are commonly used when a new predictive or diagnostic CAD-AI system has been  
1160 available in clinical practice for some time after regulatory approval [188-191]; however, care  
1161 must be taken to account for differences such as the characteristics of the patient population and  
1162 physicians' experience between the two time periods, since such differences may bias the  
1163 observed outcomes. Relevant statistical procedures such as stratification and multivariate  
1164 regression modeling can be used to account for confounding factors.

1165 The reporting of a clinical trial evaluating a CAD-AI system in the literature should allow  
1166 readers to identify potential sources of bias and, ideally, reproduce the results. Factors that may  
1167 bias or impact the results include the study population, data acquisition, characteristics of the  
1168 CAD-AI device, human-AI interaction, user training, study end-point, reference standard, and  
1169 statistical methods, all of which should be clearly identified and reported. Additionally, the  
1170 SPIRIT-AI [192] and CONSORT-AI [193] extensions provide general guidelines when drafting  
1171 clinical trial protocols or reports that target or include CAD-AI systems of any kind. It should be  
1172 noted that the CONSORT-AI statement does not yet cover advanced learning paradigms such as  
1173 continuously evolving or adaptive systems, the performance of which may change over time, and  
1174 underscore the importance of a robust post-deployment surveillance plan.

1175

### 1176 **6.5 Take Home Message on Translation to Clinic**

1177 Translation of a CAD-AI system to the clinic requires an efficient user interface, acceptance  
1178 testing to validate smooth integration into the workflow and expected performance, adequate user  
1179 training to ensure proper use and sufficient understanding of CAD-AI performance in the local

1180 clinical environment, and robust post-deployment QA procedures to monitor the consistency of  
 1181 performance over time. More advanced validation will involve prospective clinical assessments  
 1182 of the impact of CAD-AI on clinical outcomes using well-designed clinical trial protocols.

1183  
 1184

1185 **7 Discussion**

1186 The development of generalizable, robust, and reliable CAD-AI decision support systems is  
 1187 of critical importance for both laboratory proof-of-concept applications and for real-world  
 1188 applications in clinical practice.

1189 To address these important issues, the American Association of Physicists in Medicine  
 1190 (AAPM) assigned a task to the Computer-Aided Image Analysis Subcommittee (CADSC), in  
 1191 part, to develop recommendations on “best practices” for the development, performance  
 1192 assessment, and clinical translation of CAD-AI systems, which are discussed in this task group  
 1193 report. Although we focus on CAD-AI systems for medical imaging, the principles of the  
 1194 processes discussed herein are general and applicable to a broad range of AI applications in the  
 1195 medical field.

1196 A summary of the recommendations (“take home messages”), for best practices for (1) data  
 1197 collection, (2) establishing reference standards, (3) model development, (4) performance  
 1198 assessment, and (5) the translation to clinical practice is presented in Table 3.

1199  
 1200 Table 3. Summary of recommendations on the best practices and standards for the development  
 1201 and performance assessment of computer-aided decision support systems.

<b>Section</b>	<b>Take Home Message</b>
Data	In summary, proper data collection methods are of critical importance to successful training, validation, and implementation of CAD-AI algorithms. Improper collection and manipulation of data (such as improper data augmentation) can lead to an overestimation of performance or lack of generalizability.
Reference Standards	The required type and granularity of the reference standard depends on the task at hand. An objective reference standard is preferred; however, when a subjective reference standard cannot be avoided, independent assessments of multiple domain experts should be obtained and their variabilities should be evaluated.
Model Development	Training approaches, especially for deep learning algorithms, are continuously improving with the goal of achieving robust, effective, and privacy-preserving CAD-AI models. An independent test set representative of the intended use that was not employed to guide model optimization in any learning paradigm is of critical importance. Robust training methods, although important for all CAD-AI systems, are especially important for systems that may operate in clinical practice with minimal or no human supervision.
Performance Assessment	The most appropriate performance metric(s) will depend on the task and the reference standard. Often multiple performance metrics are appropriate and use of multiple metrics is frequently desirable. Power calculations should be an integral part of study design, and performance analysis should include error estimates, assessment of

	statistical significance, and preferably assessment of reproducibility.
Translation to Clinic	Translation of a CAD-AI system to the clinic requires an efficient user interface, acceptance testing to validate smooth integration into the workflow and expected performance, adequate user training to ensure proper use and sufficient understanding of CAD-AI performance in the local clinical environment, and robust post-deployment QA procedures to monitor the consistency of performance over time. More advanced validation will involve prospective clinical assessments of the impact of CAD-AI on clinical outcomes using well-designed clinical trial protocols.

1202  
1203  
1204  
1205

## Conclusions

1206 The rigor and reproducibility of CAD-AI systems will provide the foundation for the success  
1207 of such systems when translated into clinical practice. As a community, we are obligated to  
1208 ensure that the scientific integrity of systems we develop in the laboratory can endure the  
1209 variabilities and the required reliability in clinical practice to benefit patient care. The topics  
1210 discussed in this report are all essential elements of CAD-AI systems that, when diligently  
1211 considered during system development and validation, should provide the greatest opportunity  
1212 for successful clinical translation.

1213  
1214

## Disclosure Statement

1216 The members of AAPM Task Group 273 listed below disclose the following potential  
1217 Conflict(s) of Interest related to subject matter or materials presented in this document.

- 1218 Lubomir Hadjiiski - nothing to disclose
- 1219 Kenny Cha - nothing to disclose
- 1220 Heang-Ping Chan - nothing to disclose
- 1221 Karen Drukker - receives royalties from Hologic
- 1222 Lia Morra - has received funding from HealthTriage srl, not related to this work
- 1223 Janne J. Näppi - has received royalties from Hologic and from MEDIAN Technologies,  
1224 through the University of Chicago licensing, not related to this work
- 1225 Berkman Sahiner - nothing to disclose
- 1226 Hiroyuki Yoshida - has received royalties from licensing fees to Hologic and Medians  
1227 Technologies through the University of Chicago licensing, not related to this  
1228 work
- 1229 Quan Chen - has received compensations from Carina Medical LLC, not related to this work,  
1230 provides consulting services for Reflexion Medical, which is unrelated to the  
1231 content of the TG report
- 1232 Thomas M. Deserno - nothing to disclose
- 1233 Hayit Greenspan - nothing to disclose
- 1234 Henkjan Huisman - has received funding from Siemens Healthineers for a scientific research  
1235 project, not related to this work

1236 Zhimin Huo - nothing to disclose  
1237 Richard Mazurchuk - nothing to disclose  
1238 Nicholas Petrick - nothing to disclose  
1239 Daniele Regge - nothing to disclose  
1240 Ravi Samala - nothing to disclose  
1241 Ronald M. Summers - has received royalties from iCAD Medical, ScanMed, Philips, PingAn,  
1242 Translation Holdings. Lab receives research support from PingAn, not  
1243 related to this work  
1244 Kenji Suzuki - provides consulting services for Canon Medical, which is unrelated to the  
1245 content of the TG report  
1246 Georgia Tourassi - nothing to disclose  
1247 Daniel Vergara - nothing to disclose  
1248 Samuel G. Armato, III – has received royalties and licensing fees for computer-aided  
1249 diagnosis through the University of Chicago Consultant, Novartis, not  
1250 related to this work  
1251  
1252  
1253

## 1254 Acknowledgments

1255  
1256 RMS was supported in part by the Intramural Research Program of the National Institutes of  
1257 Health Clinical Center.  
1258  
1259  
1260

## 1261 References

- 1262
- 1263 1. M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero,  
1264 C. Etmann, C. McCague, L. Beer, J.R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J.H.F.  
1265 Rudd, E. Sala, C.-B. Schonlieb, and C. Aix, "Common pitfalls and recommendations for  
1266 using machine learning to detect and prognosticate for COVID-19 using chest radiographs  
1267 and CT scans," *Nature Machine Intelligence*, 3, 199-217 (2021).
  - 1268 2. M. Nagendran, Y. Chen, C.A. Lovejoy, A.C. Gordon, M. Komorowski, H. Harvey, E.J.  
1269 Topol, J.P.A. Ioannidis, G.S. Collins, and M. Maruthappu, "Artificial intelligence versus  
1270 clinicians: systematic review of design, reporting standards, and claims of deep learning  
1271 studies," *Bmj-British Medical Journal*, 368, 1-7 (2020).
  - 1272 3. R. Aggarwal, V. Sounderajah, G. Martin, D.S.W. Ting, A. Karthikesalingam, D. King, H.  
1273 Ashrafian, and A. Darzi, "Diagnostic accuracy of deep learning in medical imaging: a  
1274 systematic review and meta-analysis," *NPJ digital medicine*, 4, 65-65 (2021).
  - 1275 4. D.W. Kim, H.Y. Jang, K.W. Kim, Y. Shin, and S.H. Park, "Design Characteristics of  
1276 Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic  
1277 Analysis of Medical Images: Results from Recently Published Papers," *Korean Journal of  
1278 Radiology*, 20, 405-410 (2019).
  - 1279 5. N. Petrick, B. Sahiner, S.G. Armato, A. Bert, L. Correale, S. Delsanto, M.T. Freedman, D.  
1280 Fryd, D. Gur, L. Hadjiiski, Z.M. Huo, Y.L. Jiang, L. Morra, S. Paquerault, V. Raykar, F.

- 1281 Samuelson, R.M. Summers, G. Tourassi, H. Yoshida, B. Zheng, C. Zhou, and H.-P. Chan,  
1282 "Evaluation of computer-aided detection and diagnosis systems," *Medical Physics*, 40,  
1283 087001 (2013).
- 1284 6. Z.M. Huo, R.M. Summers, S. Paquerault, J. Lo, J. Hoffmeister, S.G. Armato, M.T.  
1285 Freedman, J. Lin, S.C.B. Lo, N. Petrick, B. Sahiner, D. Fryd, H. Yoshida, and H.-P. Chan,  
1286 "Quality assurance and training procedures for computer-aided detection and diagnosis  
1287 systems in clinical use," *Medical Physics*, 40, 077001 (2013).
- 1288 7. J.F. Cohen, D.A. Korevaar, D.G. Altman, D.E. Bruns, C.A. Gatsonis, L. Hooft, L. Irwig,  
1289 D. Levine, J.B. Reitsma, H.C.W. de Vet, and P.M.M. Bossuyt, "STARD 2015 guidelines  
1290 for reporting diagnostic accuracy studies: explanation and elaboration," *Bmj Open*, 6,  
1291 e012799 (2016).
- 1292 8. J.E. Trost, "Statistically nonrepresentative stratified sampling: A sampling technique for  
1293 qualitative studies," *Qualitative Sociology*, 9, 54-57 (1986).
- 1294 9. I. Etikan, S.A. Musa, and R.S. Alkassim, "Comparison of convenience sampling and  
1295 purposive sampling," *American journal of theoretical and applied statistics*, 5, 1-4 (2016).
- 1296 10. I. Pan, S. Agarwal, and D. Merck, "Generalizable Inter-Institutional Classification of  
1297 Abnormal Chest Radiographs Using Efficient Convolutional Neural Networks," *Journal  
1298 of Digital Imaging*, 32, 888-896 (2019).
- 1299 11. J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, and E.K. Oermann, "Variable  
1300 generalization performance of a deep learning model to detect pneumonia in chest  
1301 radiographs: A cross-sectional study," *Plos Medicine*, 15, e1002683 (2018).
- 1302 12. X. Feng, M.E. Bernard, T. Hunter, and Q. Chen, "Improving accuracy and robustness of  
1303 deep convolutional neural network based thoracic OAR segmentation," *Physics in  
1304 Medicine and Biology*, 65, 07NT01 (2020).
- 1305 13. X.X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G.  
1306 Moraes, M. Shamdas, C. Kern, J.R. Ledsam, M.K. Schmid, K. Balaskas, E.J. Topol, L.M.  
1307 Bachmann, P.A. Keane, and A.K. Denniston, "A comparison of deep learning  
1308 performance against health-care professionals in detecting diseases from medical  
1309 imaging: a systematic review and meta-analysis," *Lancet Digital Health*, 1, E271-E297  
1310 (2019).
- 1311 14. K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W.  
1312 Steyerberg, A.J. Vickers, D.F. Ransohoff, and G.S. Collins, "Transparent Reporting of a  
1313 multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD):  
1314 Explanation and Elaboration," *Annals of Internal Medicine*, 162, W1-W73 (2015).
- 1315 15. H.M. Whitney, H. Li, Y. Ji, P. Liu, and M.L. Giger, "Harmonization of radiomic features  
1316 of breast lesions across international DCE-MRI datasets," *Journal of Medical Imaging*, 7,  
1317 012707 (2020).
- 1318 16. R.M. Nishikawa, M.L. Giger, K. Doi, C.E. Metz, F.F. Yin, C.J. Vyborny, and R.A.  
1319 Schmidt, "EFFECT OF CASE SELECTION ON THE PERFORMANCE OF  
1320 COMPUTER-AIDED DETECTION SCHEMES," *Medical Physics*, 21, 265-269 (1994).
- 1321 17. R.M. Nishikawa and L.M. Yarusso, *Variations in measured performance of CAD schemes  
1322 due to database composition and scoring protocol*, in *Medical Imaging 1998: Image  
1323 Processing, Pts 1 and 2*, K.M. Hanson, Editor. 1998, p. 840-844.
- 1324 18. S.G. Armato, R.Y. Roberts, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, G. McLennan,  
1325 R.M. Engelmann, P.H. Bland, D.R. Aberle, E.A. Kazerooni, H. MacMahon, E.J.R. van  
1326 Beek, D. Yankelevitz, B.Y. Croft, and L.P. Clarke, "The lung image database consortium  
1327 (LIDC): Ensuring the integrity of expert-defined "truth", " *Academic Radiology*, 14, 1455-  
1328 1463 (2007).

- 1329 19. K.W. Clark, D.S. Gierada, G. Marquez, S.M. Moore, D.R. Maffitt, J.D. Moulton, M.A.  
1330 Wolfsberger, P. Koppel, S.R. Phillips, and F.W. Prior, "Collecting 48,000 CT Exams for  
1331 the Lung Screening Study of the National Lung Screening Trial," *Journal of Digital*  
1332 *Imaging*, 22, 667-680 (2009).
- 1333 20. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N.  
1334 Blomberg, J.W. Boiten, L.B.D. Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark,  
1335 M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran,  
1336 A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. Hoen, R. Hooft, T. Kuhn,  
1337 R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-  
1338 Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G.  
1339 Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A.  
1340 Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR Guiding  
1341 Principles for scientific data management and stewardship (vol 15, 160018, 2016),"  
1342 *Scientific Data*, 6, 6 (2019).
- 1343 21. "Summary of the HIPAA Privacy Rule," [https://www.hhs.gov/hipaa/for-](https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html)  
1344 [professional/privacy/laws-regulations/index.html](https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html),
- 1345 22. "International Compilation of Human Research Standards. 2021 Edition. Compiled by  
1346 Office for Human Research Protections, Office of the Assistant Secretary for Health, U.S.  
1347 Department of Health and Human Services "   
1348 <https://www.hhs.gov/sites/default/files/ohrp-international-compilation-2021.pdf>,
- 1349 23. H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang, and L. Floridi, "The Chinese  
1350 approach to artificial intelligence: an analysis of policy, ethics, and regulation," *Ai &*  
1351 *Society*, 36, 59-77 (2021).
- 1352 24. M. Gong, S. Wang, L. Wang, C. Liu, J. Wang, Q. Guo, H. Zheng, K. Xie, C. Wang, and  
1353 Z. Hui, "Evaluation of Privacy Risks of Patients' Data in China: Case Study," *Jmir*  
1354 *Medical Informatics*, 8, (2020).
- 1355 25. K. Pinhao and M.M. R., *Twenty reasons why GDPR compliance does not exempt*  
1356 *companies from adjusting to the LGPD*, in *International Bar Association 2021*,  
1357 <https://www.ibanet.org/article/0634B90E-98DE-40E6-953F-2F63CB481F02>.
- 1358 26. D.B. Larson, D.C. Magnus, M.P. Lungren, N.H. Shah, and C.P. Langlotz, "Ethics of  
1359 Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed  
1360 Framework," *Radiology*, 295, 675-682 (2020).
- 1361 27. J.R. Geis, A.P. Brady, C.C. Wu, J. Spencer, E. Ranschaert, J.L. Jaremko, S.G. Langer,  
1362 A.B. Kitts, J. Birch, W.F. Shields, R.V. van Genderen, E. Kotter, J.W. Gichoya, T.S.  
1363 Cook, M.B. Morgan, A. Tang, N.M. Safdar, and M. Kohli, "Ethics of Artificial  
1364 Intelligence in Radiology: Summary of the Joint European and North American  
1365 Multisociety Statement," *Journal of the American College of Radiology*, 16, 1516-1521  
1366 (2019).
- 1367 28. K.Y.E. Aryanto, M. Oudkerk, and P.M.A. van Ooijen, "Free DICOM de-identification  
1368 tools in clinical research: functioning and safety of patient privacy," *European Radiology*,  
1369 25, 3685-3695 (2015).
- 1370 29. J.D. Robinson, "Beyond the DICOM Header: Additional Issues in Deidentification,"  
1371 *American Journal of Roentgenology*, 203, W658-W664 (2014).
- 1372 30. J. Buolamwini and T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in*  
1373 *Commercial Gender Classification*, in *Proceedings of the 1st Conference on Fairness,*  
1374 *Accountability and Transparency*, A.F. Sorelle and W. Christo, Editors. 2018, PMLR:  
1375 *Proceedings of Machine Learning Research*. p. 77-91.

- 1376 31. Y. Liu, A. Jain, C. Eng, D.H. Way, K. Lee, P. Bui, K. Kanada, G.D. Marinho, J. Gallegos,  
1377 S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G.S. Corrado,  
1378 L.H. Peng, D.R. Webster, D. Ai, S.J. Huang, Y. Liu, R.C. Dunn, and D. Coz, "A deep  
1379 learning system for differential diagnosis of skin diseases," *Nature Medicine*, 26, 900-908  
1380 (2020).
- 1381 32. J.W. Gichoya, I. Banerjee, A.R. Bhimireddy, J.L. Burns, L.A. Celi, L.C. Chen, R. Correa,  
1382 N. Dullerud, M. Ghassemi, S.C. Huang, P.C. Kuo, M.P. Lungren, L.J. Palmer, B.J. Price,  
1383 S. Purkayastha, A.T. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H.  
1384 Trivedi, R.Y. Wang, Z. Zaiman, and H.R. Zhang, "AI recognition of patient race in  
1385 medical imaging: a modelling study," *Lancet Digital Health*, 4, E406-E414 (2022).
- 1386 33. S. Shrestha and S. Das, "Exploring gender biases in ML and AI academic research  
1387 through systematic literature review," *Frontiers in artificial intelligence*, 5, 976838-  
1388 976838 (2022).
- 1389 34. I. Dankwa-Mullan and D. Weeraratne, "Artificial Intelligence and Machine Learning  
1390 Technologies in Cancer Care: Addressing Disparities, Bias, and Data Diversity," *Cancer  
1391 Discovery*, 12, 1423-1427 (2022).
- 1392 35. H.-P. Chan, S.C.B. Lo, B. Sahiner, K.L. Lam, and M.A. Helvie, "Computer-aided  
1393 detection of mammographic microcalcifications: Pattern recognition with an artificial  
1394 neural network," *Medical Physics*, 22, 1555-1567 (1995).
- 1395 36. A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep  
1396 convolutional neural networks," in *Advances in Neural Information Processing Systems*,  
1397 pp. 1097-1105 (2012).
- 1398 37. I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.  
1399 Courville, and Y. Bengio, "Generative Adversarial Nets," arXiv:1406.2661v1 (2014).
- 1400 38. M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "SYNTHETIC  
1401 DATA AUGMENTATION USING GAN FOR IMPROVED LIVER LESION  
1402 CLASSIFICATION," in *15th IEEE International Symposium on Biomedical Imaging  
1403 (ISBI)*, Washington, DC. pp. 289-293 (2018).
- 1404 39. K.H. Cha, N. Petrick, A. Pezeshk, C.G. Graff, D. Sharma, A. Badal, and B. Sahiner,  
1405 "Evaluation of data augmentation via synthetic images for improved breast mass detection  
1406 on mammograms using deep learning," *Journal of medical imaging (Bellingham, Wash.)*,  
1407 7, 012703-012703 (2020).
- 1408 40. A. Hagiwara, S. Fujita, Y. Ohno, and S. Aoki, "Variability and Standardization of  
1409 Quantitative Imaging Monoparametric to Multiparametric Quantification, Radiomics, and  
1410 Artificial Intelligence," *Investigative Radiology*, 55, 601-616 (2020).
- 1411 41. B. Graham, "Kaggle diabetic retinopathy detection competition report," University of  
1412 Warwick, (2015).
- 1413 42. K. Robinson, H. Li, L. Lan, D. Schacht, and M. Giger, "Radiomics robustness assessment  
1414 and classification evaluation: A two-stage method demonstrated on multivendor FFDM,"  
1415 *Medical Physics*, 46, 2145-2156 (2019).
- 1416 43. B. Baessler, K. Weiss, and D.P. dos Santos, "Robustness and Reproducibility of  
1417 Radiomics in Magnetic Resonance Imaging A Phantom Study," *Investigative Radiology*,  
1418 54, 221-228 (2019).
- 1419 44. S.A. Mali, A. Ibrahim, H.C. Woodruff, V. Andrearczyk, H. Mueller, S. Primakov, Z.  
1420 Salahuddin, A. Chatterjee, and P. Lambin, "Making Radiomics More Reproducible across  
1421 Scanner and Imaging Protocol Variations: A Review of Harmonization Methods," *Journal  
1422 of Personalized Medicine*, 11, (2021).

- 1423 45. L. Gallardo-Estrella, D.A. Lynch, M. Prokop, D. Stinson, J. Zach, P.F. Judy, B. van  
1424 Ginneken, and E.M. van Rikxoort, "Normalizing computed tomography data  
1425 reconstructed with different filter kernels: effect on emphysema quantification," *European*  
1426 *Radiology*, 26, 478-486 (2016).
- 1427 46. M. Liu, P. Maiti, S. Thomopoulos, A. Zhu, Y. Chai, H. Kim, and N. Jahanshad, "Style  
1428 Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization," in  
1429 *International Conference on Medical Image Computing and Computer Assisted*  
1430 *Intervention (MICCAI)*, Electr Network. pp. 313-322 (2021).
- 1431 47. R. Rai, L.C. Holloway, C. Brink, M. Field, R.L. Christiansen, Y. Sun, M.B. Barton, and  
1432 G.P. Liney, "Multicenter evaluation of MRI-based radiomic features: A phantom study,"  
1433 *Medical Physics*, 47, 3054-3063 (2020).
- 1434 48. J.-P. Fortin, D. Parker, B. Tunc, T. Watanabe, M.A. Elliott, K. Ruparel, D.R. Roalf, T.D.  
1435 Satterthwaite, R.C. Gur, R.E. Gur, R.T. Schultz, R. Verma, and R.T. Shinohara,  
1436 "Harmonization of multi-site diffusion tensor imaging data," *Neuroimage*, 161, 149-170  
1437 (2017).
- 1438 49. F. Orlhac, F. Frouin, C. Nioche, N. Ayache, and I. Buvat, "Validation of a Method to  
1439 Compensate Multicenter Effects Affecting CT Radiomics," *Radiology*, 291, 52-58 (2019).
- 1440 50. T. Nakahara, H. Daisaki, Y. Yamamoto, T. Iimori, K. Miyagawa, T. Okamoto, Y. Owaki,  
1441 N. Yada, K. Sawada, R. Tokorodani, and M. Jinzaki, "Use of a digital phantom developed  
1442 by QIBA for harmonizing SUVs obtained from the state-of-the-art SPECT/CT systems: a  
1443 multicenter study," *Ejnmri Research*, 7, (2017).
- 1444 51. H. Keller, T. Shek, B. Driscoll, Y. Xu, B. Nghiem, S. Nehmeh, M. Grkovski, C.R.  
1445 Schmidlein, M. Budzevich, Y. Balagurunathan, J.J. Sunderland, R.R. Beichel, C. Uribe,  
1446 T.-Y. Lee, F. Li, D.A. Jaffray, and I. Yeung, "Noise-Based Image Harmonization  
1447 Significantly Increases Repeatability and Reproducibility of Radiomics Features in PET  
1448 Images: A Phantom Study," *Tomography*, 8, 1113-1128 (2022).
- 1449 52. G. Revesz, H.L. Kundel, and M. Bonitatibus, "THE EFFECT OF VERIFICATION ON  
1450 THE ASSESSMENT OF IMAGING TECHNIQUES," *Investigative Radiology*, 18, 194-  
1451 198 (1983).
- 1452 53. D.P. Miller, K.F. O'Shaughnessy, S.A. Wood, and R.A. Castellino, *Gold standards and*  
1453 *expert panels: A pulmonary nodule case study with challenges and solutions*, in *Medical*  
1454 *Imaging 2004: Image Perception, Observer Performance, and Technology Assessment*,  
1455 D.P. Chakraborty and M.P. Eckstein, Editors. 2004, p. 173-184.
- 1456 54. Y. Jiang, "A Monte Carlo simulation method to understand expert-panel consensus truth  
1457 and double readings," *Medical Image Perception Conference XII. 2007. The University of*  
1458 *Iowa, Iowa City, IA, (2007).*
- 1459 55. S.G. Armato, R.Y. Roberts, M. Kocherginsky, D.R. Aberle, E.A. Kazerooni, H.  
1460 MacMahon, E.J.R. van Beek, D. Yankelevitz, G. McLennan, M.F. McNitt-Gray, C.R.  
1461 Meyer, A.P. Reeves, P. Caligiuri, L.E. Quint, B. Sundaram, B.Y. Croft, and L.P. Clarke,  
1462 "Assessment of Radiologist Performance in the Detection of Lung Nodules: Dependence  
1463 on the Definition of "Truth"," *Academic Radiology*, 16, 28-38 (2009).
- 1464 56. C. Zhou, H.-P. Chan, A. Chughtai, S. Patel, J. Kuriakose, L.M. Hadjiiski, J. Wei, and E.A.  
1465 Kazerooni, "Variabilities in Reference Standard by Radiologists and Performance  
1466 Assessment in Detection of Pulmonary Embolism in CT Pulmonary Angiography,"  
1467 *Journal of Digital Imaging*, 32, 1089-1096 (2019).
- 1468 57. B. Sahiner, H.-P. Chan, L.M. Hadjiiski, P.N. Cascade, E.A. Kazerooni, A.R. Chughtai, C.  
1469 Poopat, T. Song, L. Frank, J. Stojanovska, and A. Attili, "Effect of CAD on radiologists'

- 1470 detection of lung nodules on thoracic CT scans: Analysis of an observer performance  
1471 study by nodule size," *Academic Radiology*, 1518-1530 (2009).
- 1472 58. A. Wenzel and H. Hintze, "The choice of gold standard for evaluating tests for caries  
1473 diagnosis," *Dentomaxillofacial Radiology*, 28, 132-136 (1999).
- 1474 59. T.M. Lehmann, *From plastic to gold: A unified classification scheme for reference  
1475 standards in medical image processing*, in *Medical Imaging 2002: Image Processing, Vol  
1476 1-3*, M. Sonka and J.M. Fitzpatrick, Editors. 2002, p. 1819-1827.
- 1477 60. F. Li, R. Engelmann, S.G. Armato, and H. MacMahon, "Computer-Aided Nodule  
1478 Detection System: Results in an Unselected Series of Consecutive Chest Radiographs,"  
1479 *Academic Radiology*, 22, 475-480 (2015).
- 1480 61. D.F. Yankelevitz and C.I. Henschke, "Does 2-year stability imply that pulmonary nodules  
1481 are benign?," *American Journal of Roentgenology*, 168, 325-328 (1997).
- 1482 62. G.J.S. Litjens, J.O. Barentsz, N. Karssemeijer, and H.J. Huisman, "Clinical evaluation of  
1483 a computer-aided diagnosis system for determining cancer aggressiveness in prostate  
1484 MRI," *European Radiology*, 25, 3187-3199 (2015).
- 1485 63. "DREAM. The digital mammography dream challenge.  
1486 [https://www.synapse.org/Digital\\_Mammography\\_DREAM\\_challenge](https://www.synapse.org/Digital_Mammography_DREAM_challenge)," (2017).
- 1487 64. S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T.  
1488 Back, M. Chesus, G.C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F.J. Gilbert,  
1489 M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C.J. Kelly, D. King, J.R.  
1490 Ledsam, D. Melnick, H. Mostofi, L. Peng, J.J. Reicher, B. Romera-Paredes, R.  
1491 Sidebottom, M. Suleyman, D. Tse, K.C. Young, J. De Fauw, and S. Shetty, "International  
1492 evaluation of an AI system for breast cancer screening," *Nature*, 577, 89-94 (2020).
- 1493 65. C.R. Meyer, T.D. Johnson, G. McLennan, D.R. Aberle, E.A. Kazerooni, H. MacMahon,  
1494 B.F. Mullan, D.F. Yankelevitz, E.J.R. van Beek, S.G. Armato, M.F. McNitt-Gray, A.P.  
1495 Reeves, D. Gur, C.I. Henschke, E.A. Hoffman, P.H. Bland, G. Laderach, R. Pais, D.  
1496 Qing, C. Piker, J.F. Guo, A. Starkey, D. Max, B.Y. Croft, and L.P. Clarke, "Evaluation of  
1497 lung MDCT nodule annotation across radiologists and methods," *Academic Radiology*,  
1498 13, 1254-1265 (2006).
- 1499 66. J. Tan, J. Pu, B. Zheng, X. Wang, and J.K. Leader, "Computerized comprehensive data  
1500 analysis of Lung Imaging Database Consortium (LIDC)," *Medical Physics*, 37, 3802-  
1501 3808 (2010).
- 1502 67. K. Yan, X. Wang, L. Lu, and R.M. Summers, "DeepLesion: automated mining of large-  
1503 scale lesion annotations and universal lesion detection with deep learning," *Journal of  
1504 Medical Imaging*, 5, 036501 (2018).
- 1505 68. L. Oakden-Rayner, "Exploring Large-scale Public Medical Image Datasets," *Academic  
1506 Radiology*, 27, 106-112 (2020).
- 1507 69. D.A. Bluemke, L. Moy, M.A. Bredella, B.B. Ertl-Wagner, K.J. Fowler, V.J. Goh, E.F.  
1508 Halpern, C.P. Hess, M.L. Schiebler, and C.R. Weiss, "Assessing Radiology Research on  
1509 Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers-From the  
1510 Radiology Editorial Board," *Radiology*, 294, 487-489 (2020).
- 1511 70. S. Goel, Y. Sharma, M.-L. Jauer, and T.M. Deserno, "WeLineation: Crowdsourcing  
1512 delineations for reliable ground truth estimation," *Medical Imaging 2020: Imaging  
1513 Informatics for Healthcare, Research, and Applications*, 11318, 113180C (2020).
- 1514 71. T.B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, J.E. Burns, and  
1515 R.M. Summers, "Distributed Human Intelligence for Colonic Polyp Classification in  
1516 Computer-aided Detection for CT Colonography," *Radiology*, 262, 824-833 (2012).

- 1517 72. M.-L. Jauer, S. Goel, Y. Sharma, T.M. Deserno, M. Gijs, T.T.J.M. Berendschot, C.J.F.  
1518 Bertens, and R.M.M.A. Nuijts, "STAPLE performance assessed on crowdsourced sclera  
1519 segmentations," *Medical Imaging 2020: Imaging Informatics for Healthcare, Research,  
1520 and Applications*, 11318, 113180K (2020).
- 1521 73. A. Badano, C.G. Graff, A. Badal, D. Sharma, R. Zeng, F.W. Samuelson, S.J. Glick, and  
1522 K.J. Myers, "Evaluation of Digital Breast Tomosynthesis as Replacement of Full-Field  
1523 Digital Mammography Using an In Silico Imaging Trial," *JAMA Network Open*, 1,  
1524 e185474-e185474 (2018).
- 1525 74. E. Abadi, W.P. Segars, H. Chalian, and E. Samei, "Virtual Imaging Trials for Coronavirus  
1526 Disease (COVID-19)," *American Journal of Roentgenology*, 216, 362-368 (2021).
- 1527 75. R.K. Samala, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, and C.D. Richter, "Generalization  
1528 error analysis for deep convolutional neural network with transfer learning in breast  
1529 cancer diagnosis," *Physics in Medicine and Biology*, 65, 105002 (2020).
- 1530 76. M. Rajchl, M.C.H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M.  
1531 Damodaram, M.A. Rutherford, J.V. Hajnal, B. Kainz, and D. Rueckert, "DeepCut: Object  
1532 Segmentation From Bounding Box Annotations Using Convolutional Neural Networks,"  
1533 *Ieee Transactions on Medical Imaging*, 36, 674-683 (2017).
- 1534 77. S.K. Warfield, K.H. Zou, and W.M. Wells, "Simultaneous truth and performance level  
1535 estimation (STAPLE): An algorithm for the validation of image segmentation," *Ieee  
1536 Transactions on Medical Imaging*, 23, 903-921 (2004).
- 1537 78. N. Petrick, B. Sahiner, H.-P. Chan, M.A. Helvie, S. Paquerault, and L.M. Hadjiiski,  
1538 "Breast cancer detection: Evaluation of a mass-detection algorithm for computer-aided  
1539 diagnosis - Experience in 263 patients," *Radiology*, 224, 217-224 (2002).
- 1540 79. M. Kallergi, G.M. Carney, and J. Gaviria, "Evaluating the performance of detection  
1541 algorithms in digital mammography," *Medical Physics*, 26, 267-275 (1999).
- 1542 80. S.K. Zhou, H. Greenspan, C. Davatzikos, J.S. Duncan, B. Van Ginneken, A. Madabhushi,  
1543 J.L. Prince, D. Rueckert, and R.M. Summers, "A Review of Deep Learning in Medical  
1544 Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and  
1545 Future Promises," *Proceedings of the Ieee*, 109, 820-838 (2021).
- 1546 81. D. Gur, R.F. Wagner, and H.-P. Chan, "On the repeated use of databases for testing  
1547 incremental improvement of computer-aided detection schemes," *Academic Radiology*,  
1548 11, 103-105 (2004).
- 1549 82. K. Fukunaga, "Introduction to statistical pattern recognition," 2nd edition. Academic  
1550 Press, San Diego, (1990).
- 1551 83. Q. Li and K. Doi, "Comparison of typical evaluation methods for computer-aided  
1552 diagnostic schemes: Monte Carlo simulation study," *Medical Physics*, 34, 871-876  
1553 (2007).
- 1554 84. B. Efron, "ESTIMATING THE ERROR RATE OF A PREDICTION RULE -  
1555 IMPROVEMENT ON CROSS-VALIDATION," *Journal of the American Statistical  
1556 Association*, 78, 316-331 (1983).
- 1557 85. B. Sahiner, H.-P. Chan, and L. Hadjiiski, "Classifier performance prediction for  
1558 computer-aided diagnosis using a limited dataset," *Medical Physics*, 35, 1559-1570  
1559 (2008).
- 1560 86. J.M. Bland and D.G. Altman, "Statistics Notes: Bootstrap resampling methods," *Bmj-  
1561 British Medical Journal*, 350, h2622 (2015).
- 1562 87. R.K. Samala, H.P. Chan, L. Hadjiiski, and M.A. Helvie, "Risks of feature leakage and  
1563 sample size dependencies in deep feature extraction for breast mass classification,"  
1564 *Medical Physics*, 48, 2827-2837 (2021).

- 1565 88. S. Russell and P. Norving, "Artificial intelligence: a modern approach," 4th Edition,  
1566 Pearson, USA, (2020).
- 1567 89. C. Bishop, "Pattern recognition and machine learning," Springer, Singapore, (2006).
- 1568 90. P. Winston, "Artificial Intelligence," 3rd Edition, Addison-Wesley, USA, (1993).
- 1569 91. A. Jaiswal, A.R. Babu, M.Z. Zadeh, D. Banerjee, and F. Makedon, "A Survey on  
1570 Contrastive Self-Supervised Learning," *Technologies*, 9, (2021).
- 1571 92. J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, and T. Cai, "Multi -task contrastive  
1572 learning for automatic CT and X-ray diagnosis of COVID-19," *Pattern Recognition*, 114,  
1573 (2021).
- 1574 93. J.J. Nappi, R. Tachibana, T. Hironaka, and H. Yoshida, "Electronic cleansing by unpaired  
1575 contrastive learning in non-cathartic laxative-free CT colonography," *Proc SPIE Medical  
1576 Imaging*, 12037, 120370S (2022).
- 1577 94. N. Tajbakhsh, Y.F. Hu, J.L. Cao, X.J. Yan, Y. Xiao, Y. Lu, J.M. Liang, D. Terzopoulos,  
1578 and X.W. Ding, *Surrogate Supervision For Medical Image Analysis: Effective Deep  
1579 Learning From Limited Quantities of Labeled Data*, in *2019 Ieee 16th International  
1580 Symposium on Biomedical Imaging*. 2019, p. 1251-1255.
- 1581 95. R. Tachibana, J.J. Nappi, T. Hironaka, and H. Yoshida, "Self-Supervised adversarial  
1582 learning with a limited dataset for electronic cleansing in computed tomographic  
1583 colonography: a preliminary feasibility study," *Cancers*, 14, 4125 (2022).
- 1584 96. Z. Zhou, V. Sodha, M.M.R. Siddiquee, R. Feng, N. Tajbakhsh, M.B. Gotway, and J.  
1585 Liang, *Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis*, in  
1586 *Medical Image Computing and Computer Assisted Intervention - Miccai 2019, Pt Iv*, D.  
1587 Shen, *et al.*, Editors. 2019, p. 384-393.
- 1588 97. S. Beiden, G. Campbell, K. Meier, and R. Wagner, "The problem of ROC analysis  
1589 without truth: the EM algorithm and the information matrix," *Proc SPIE Medical  
1590 Imaging*, 3981, 126-134 (2000).
- 1591 98. V. Cheplygina, M. de Bruijne, and J.P.W. Pluim, "Not-so-supervised: A survey of semi-  
1592 supervised, multi-instance, and transfer learning in medical image analysis," *Medical  
1593 Image Analysis*, 54, 280-296 (2019).
- 1594 99. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 521, 436-444 (2015).
- 1595 100. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, *How transferable are features in deep  
1596 neural networks ?*, in *Advances in Neural Information Processing Systems*, Z.  
1597 Ghahramani, *et al.*, Editors. 2014, p. 3320-3328.
- 1598 101. Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, *CHEST  
1599 PATHOLOGY DETECTION USING DEEP LEARNING WITH NON-MEDICAL  
1600 TRAINING*, in *2015 Ieee 12th International Symposium on Biomedical Imaging*. 2015, p.  
1601 294-297.
- 1602 102. H.C. Shin, H.R. Roth, M.C. Gao, L. Lu, Z.Y. Xu, I. Nogues, J.H. Yao, D. Mollura, and  
1603 R.M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection:  
1604 CNN Architectures, Dataset Characteristics and Transfer Learning," *Ieee Transactions on  
1605 Medical Imaging*, 35, 1285-1298 (2016).
- 1606 103. I. Diamant, Y. Bar, O. Geva, L. Wolf, G. Zimmerman, S. Lieberman, E. Konen, and H.  
1607 Greenspan, *Chest Radiograph Pathology Categorization via Transfer Learning*. Deep  
1608 Learning for Medical Image Analysis, eds. S.K. Zhou, H. Greenspan, and D. Shen. 2017.  
1609 299-320.
- 1610 104. R.K. Samala, H.-P. Chan, L. Hadjiiski, M.A. Helvie, J. Wei, and K. Cha, "Mass detection  
1611 in digital breast tomosynthesis: Deep convolutional neural network with transfer learning  
1612 from mammography," *Medical Physics*, 43, 6654-6666 (2016).

- 1613 105. N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, and J.M.  
1614 Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or  
1615 Fine Tuning?," *Ieee Transactions on Medical Imaging*, 35, 1299-1312 (2016).
- 1616 106. J. Yang, X. Huang, Y. He, J. Xu, C. Yang, G. Xu, and B. Ni, "Reinventing 2D  
1617 Convolutions for 3D Images," *Ieee Journal of Biomedical and Health Informatics*, 25,  
1618 3009-3018 (2021).
- 1619 107. R. Tachibana, J.J. Nappi, J. Ota, N. Kohlhase, T. Hironaka, S.H. Kim, D. Regge, and H.  
1620 Yoshida, "Deep Learning Electronic Cleansing for Single- and Dual-Energy CT  
1621 Colonography," *Radiographics*, 38, 2034-2050 (2018).
- 1622 108. R.K. Samala, H.-P. Chan, L. Hadjiiski, M.A. Helvie, C.D. Richter, and K.H. Cha, "Breast  
1623 Cancer Diagnosis in Digital Breast Tomosynthesis: Effects of Training Sample Size on  
1624 Multi-Stage Transfer Learning Using Deep Neural Nets," *Ieee Transactions on Medical  
1625 Imaging*, 38, 686-696 (2019).
- 1626 109. X. Mei, Z. Liu, P.M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K.E.  
1627 Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z.A. Fayad, and Y. Yang,  
1628 "RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective  
1629 Transfer Learning," *Radiology: Artificial Intelligence* 4, e210315 (2022).
- 1630 110. M. Heker and H. Greenspan, "Joint Liver Lesion Segmentation and Classification via  
1631 Transfer Learning," *arXiv preprint arXiv:2004.12352*, (2020).
- 1632 111. R. Caruana, "Multitask learning," In: *Learning to learn*. Springer, 95-133 (1998).
- 1633 112. R.K. Samala, H.-P. Chan, L.M. Hadjiiski, M.A. Helvie, K.H. Cha, and C.D. Richter,  
1634 "Multi-task transfer learning deep convolutional neural network: application to computer-  
1635 aided diagnosis of breast cancer on mammograms," *Physics in Medicine and Biology*, 62,  
1636 8894-8908 (2017).
- 1637 113. J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, *Dataset Shift  
1638 in Machine Learning*. ACM Digital Library, 2009, The MIT Press. 1-248.
- 1639 114. D.C. Castro, I. Walker, and G. B., "Causality matters in medical imaging," *arXiv preprint  
1640 arXiv:1912.08142*, (2019).
- 1641 115. G. Csurka, *A Comprehensive Survey on Domain Adaptation for Visual Applications*, in  
1642 *Domain Adaptation in Computer Vision Applications*, G. Csurka, Editor. 2017, Springer:  
1643 p. 1-35.
- 1644 116. K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon,  
1645 A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, *Unsupervised Domain Adaptation in  
1646 Brain Lesion Segmentation with Adversarial Networks*, in *Information Processing in  
1647 Medical Imaging*, M. Niethammer, *et al.*, Editors. 2017, p. 597-609.
- 1648 117. A.F. Frangi, S.A. Tsaftaris, and J.L. Prince, "Simulation and Synthesis in Medical  
1649 Imaging," *Ieee Transactions on Medical Imaging*, 37, 673-679 (2018).
- 1650 118. F. Mahmood, R. Chen, and N.J. Durr, "Unsupervised Reverse Domain Adaptation for  
1651 Synthetic Medical Images via Adversarial Training," *Ieee Transactions on Medical  
1652 Imaging*, 37, 2572-2581 (2018).
- 1653 119. H.C. Shin, N.A. Tenenholtz, J.K. Rogers, C.G. Schwarz, M.L. Senjem, J.L. Gunter, K.P.  
1654 Andriole, and M. Michalski, *Medical Image Synthesis for Data Augmentation and  
1655 Anonymization Using Generative Adversarial Networks*, in *Simulation and Synthesis in  
1656 Medical Imaging*, A. Gooya, *et al.*, Editors. 2018, p. 1-11.
- 1657 120. V. Sandfort, K. Yan, P.J. Pickhardt, and R.M. Summers, "Data augmentation using  
1658 generative adversarial networks (CycleGAN) to improve generalizability in CT  
1659 segmentation tasks," *Scientific Reports*, 9, 16884 (2019).

- 1660 121. H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A.Y. Arcas, *Communication-*  
1661 *Efficient Learning of Deep Networks from Decentralized Data*, in *Artificial Intelligence*  
1662 *and Statistics, Vol 54*, A. Singh and J. Zhu, Editors. 2017, p. 1273-1282.
- 1663 122. J. Konecny, H.B. McMahan, F.X. Yu, P. Richtarik, A.T. Suresh, and D. Bacon,  
1664 "Federated learning: Strategies for improving communication efficiency," arXiv preprint  
1665 arXiv:1610.05492, (2016).
- 1666 123. K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D.L. Rubin, and  
1667 J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical  
1668 imaging," *Journal of the American Medical Informatics Association*, 25, 945-954 (2018).
- 1669 124. N. Rieke, J. Hancox, W. Li, F. Milletari, H.R. Roth, S. Albarqouni, S. Bakas, M.N.  
1670 Galtier, B.A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R.M. Summers, A. Trask,  
1671 D. Xu, M. Baust, and M.J. Cardoso, "The future of digital health with federated learning,"  
1672 *Npj Digital Medicine*, 3, 119 (2020).
- 1673 125. X. Li, Y. Gu, N. Dvornek, L.H. Staib, P. Ventola, and J.S. Duncan, "Multi-site fMRI  
1674 analysis using privacy-preserving federated learning and domain adaptation: ABIDE  
1675 results," *Medical Image Analysis*, 65, 101765 (2020).
- 1676 126. P. McClure, C.Y. Zheng, J.R. Kaczmarzyk, J.A. Lee, S.S. Ghosh, D. Nielson, P.  
1677 Bandettini, and F. Pereira, *Distributed Weight Consolidation: A Brain Segmentation Case*  
1678 *Study*, in *Advances in Neural Information Processing Systems 31*, S. Bengio, et al.,  
1679 Editors. 2018, p. 4093-4103.
- 1680 127. P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, , and R.G.  
1681 d'Oliveira, "Advances and open problems in federated learning," arXiv preprint  
1682 arXiv:1912.04977, (2019).
- 1683 128. S. Grossberg, "Adaptive Resonance Theory: How a brain learns to consciously attend,  
1684 learn, and recognize a changing world," *Neural Networks*, 37, 1-47 (2013).
- 1685 129. G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning  
1686 with neural networks: A review," *Neural Networks*, 113, 54-71 (2019).
- 1687 130. R.M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive*  
1688 *Sciences*, 3, 128-135 (1999).
- 1689 131. I.J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical  
1690 investigation of catastrophic forgetting in gradient-based neural networks," arXiv preprint  
1691 arXiv:1312.6211, (2013).
- 1692 132. C.E. Metz, "ROC METHODOLOGY IN RADIOLOGIC IMAGING," *Investigative*  
1693 *Radiology*, 21, 720-733 (1986).
- 1694 133. D.P. Chakraborty and L.H.L. Winter, "FREE-RESPONSE METHODOLOGY -  
1695 ALTERNATE ANALYSIS AND A NEW OBSERVER-PERFORMANCE  
1696 EXPERIMENT," *Radiology*, 174, 873-881 (1990).
- 1697 134. B.D. Gallas, H.-P. Chan, C.J. D'Orsi, L.E. Dodd, M.L. Giger, D. Gur, E.A. Krupinski,  
1698 C.E. Metz, K.J. Myers, N.A. Obuchowski, B. Sahiner, A.Y. Toledano, and M.L. Zuley,  
1699 "Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA,"  
1700 *Academic Radiology*, 19, 463-477 (2012).
- 1701 135. K. Doi, H. MacMahon, S. Katsuragawa, R.M. Nishikawa, and Y.L. Jiang, "Computer-  
1702 aided diagnosis in radiology: potential and pitfalls," *European Journal of Radiology*, 31,  
1703 97-109 (1999).
- 1704 136. F.E. Harrell, R.M. Califf, D.B. Pryor, K.L. Lee, and R.A. Rosati, "EVALUATING THE  
1705 YIELD OF MEDICAL TESTS," *Jama-Journal of the American Medical Association*,  
1706 247, 2543-2546 (1982).

- 1707 137. F.E. Harrell, K.L. Lee, and D.B. Mark, "Multivariable prognostic models: Issues in  
1708 developing models, evaluating assumptions and adequacy, and measuring and reducing  
1709 errors," *Statistics in Medicine*, 15, 361-387 (1996).
- 1710 138. N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its  
1711 consideration," *Cancer Chemotherap Rep*, 50, 163-170 (1966).
- 1712 139. P. Therasse, S.G. Arbuck, E.A. Eisenhauer, J. Wanders, R.S. Kaplan, L. Rubinstein, J.  
1713 Verweij, M. Van Glabbeke, A.T. van Oosterom, M.C. Christian, and S.G. Gwyther, "New  
1714 Guidelines to Evaluate the Response to Treatment in Solid Tumors," *JNCI: Journal of the  
1715 National Cancer Institute*, 92, 205-216 (2000).
- 1716 140. E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey,  
1717 S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D.  
1718 Lacombe, and J. Verweij, "New response evaluation criteria in solid tumours: Revised  
1719 RECIST guideline (version 1.1)," *European Journal of Cancer*, 45, 228-247 (2009).
- 1720 141. P.R. Cohen, "Empirical methods for artificial intelligence," The MIT Press, (1995).
- 1721 142. X.-H. Zhou, N.A. Obuchowski, and D.K. McClish, "Statistical methods in diagnostic  
1722 medicine. Wiley; New York," (2002).
- 1723 143. P. Schober, S.M. Bossers, and L.A. Schwarte, "Statistical Significance Versus Clinical  
1724 Importance of Observed Effect Sizes: What Do P Values and Confidence Intervals Really  
1725 Represent?," *Anesthesia and Analgesia*, 126, 1068-1072 (2018).
- 1726 144. S.N. Goodman, "Toward evidence-based medical statistics. 1: The P value fallacy,"  
1727 *Annals of Internal Medicine*, 130, 995-1004 (1999).
- 1728 145. M. Aickin and H. Gensler, "Adjusting for multiple testing when reporting research  
1729 results: The Bonferroni vs Holm methods," *American Journal of Public Health*, 86, 726-  
1730 728 (1996).
- 1731 146. P. Rajpurkar, J. Irvin, R.L. Ball, K.L. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A.  
1732 Bagul, C.P. Langlotz, B.N. Patel, K.W. Yeom, K. Shpanskaya, F.G. Blankenberg, J.  
1733 Seekins, T.J. Amrhein, D.A. Mong, S.S. Halabi, E.J. Zucker, A.Y. Ng, and M.P. Lungren,  
1734 "Deep learning for chest radiograph diagnosis: A retrospective comparison of the  
1735 CheXNeXt algorithm to practicing radiologists," *Plos Medicine*, 15, e1002686 (2018).
- 1736 147. A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, and S. Thrun,  
1737 "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*,  
1738 542, 115-118 (2017).
- 1739 148. A. Rodriguez-Ruiz, K. Lang, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser,  
1740 T.H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M.G. Wallis, I. Andersson, S.  
1741 Zackrisson, R.M. Mann, and I. Sechopoulos, "Stand-Alone Artificial Intelligence for  
1742 Breast Cancer Detection in Mammography: Comparison With 101 Radiologists," *Jnci-  
1743 Journal of the National Cancer Institute*, 111, 916-922 (2019).
- 1744 149. H.P. Chan, K. Doi, C.J. Vyborny, R.A. Schmidt, C.E. Metz, K.L. Lam, T. Ogura, Y.Z.  
1745 Wu, and H. Macmahon, "IMPROVEMENT IN RADIOLOGISTS DETECTION OF  
1746 CLUSTERED MICROCALCIFICATIONS ON MAMMOGRAMS - THE POTENTIAL  
1747 OF COMPUTER-AIDED DIAGNOSIS," *Investigative Radiology*, 25, 1102-1110 (1990).
- 1748 150. L.M. Hadjiiski, H.-P. Chan, B. Sahiner, M.A. Helvie, M. Roubidoux, C. Blane, C.  
1749 Paramagul, N. Petrick, J. Bailey, K. Klein, M. Foster, S. Patterson, D. Adler, A. Nees, and  
1750 J. Shen, "Breast Masses: Computer-aided Diagnosis with Serial Mammograms,"  
1751 *Radiology*, 240, 343-356 (2006).
- 1752 151. S.V. Beiden, R.F. Wagner, K. Doi, R.M. Nishikawa, M. Freedman, S.C. Ben Lo, and  
1753 X.W. Xu, "Independent versus sequential reading in ROC studies of computer-assist

- 1754 modalities: Analysis of components of variance," *Academic Radiology*, 9, 1036-1043  
1755 (2002).
- 1756 152. C.E. Metz, "Some practical issues of experimental design and data analysis in radiological  
1757 ROC studies," *Investigative Radiology*, 24, 234-245 (1989).
- 1758 153. "U.S. Food and Drug Administration. Guidance for industry and FDA staff: Computer-  
1759 assisted detection devices applied to radiology images and radiology device data –  
1760 premarket notification [510(k)] submissions. 2012 Nov. 21, 2017]; , " Available from:  
1761 [http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/Guidance](http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM187294.pdf)  
1762 [Documents/UCM187294.pdf](http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM187294.pdf), (2017).
- 1763 154. "U.S. Food and Drug Administration. Guidance for industry and FDA staff: Clinical  
1764 performance assessment: Considerations for computer-assisted detection devices applied  
1765 to radiology images and radiology device data - premarket approval (PMA) and  
1766 premarket notification [510(k)] submissions. 2012 Nov. 21, 2017]; Available from:  
1767 [http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/Guidance](http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM187315.pdf)  
1768 [Documents/UCM187315.pdf](http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM187315.pdf)," (2017).
- 1769 155. F.W. Samuelson and C.K. Abbey, "The Reproducibility of Changes in Diagnostic Figures  
1770 of Merit Across Laboratory and Clinical Imaging Reader Studies," *Academic Radiology*,  
1771 24, 1436-1446 (2017).
- 1772 156. B.D. Gallas, W. Chen, E. Cole, R. Ochs, N. Petrick, E.D. Pisano, B. Sahiner, F.W.  
1773 Samuelson, and K.J. Myers, "Impact of prevalence and case distribution in lab-based  
1774 diagnostic imaging studies," *Journal of Medical Imaging*, 6, 015501 (2019).
- 1775 157. R.F. Wagner, C.E. Metz, and G. Campbell, "Assessment of medical imaging systems and  
1776 computer aids: A tutorial review," *Academic Radiology*, 14, 723-748 (2007).
- 1777 158. N.A. Obuchowski, "New methodological tools for multiple-reader ROC studies,"  
1778 *Radiology*, 243, 10-12 (2007).
- 1779 159. H.-P. Chan, B. Sahiner, R.F. Wagner, and N. Petrick, "Classifier design for computer-  
1780 aided diagnosis: Effects of finite sample size on the mean performance of classical and  
1781 neural network classifiers," *Medical Physics*, 26, 2654-2668 (1999).
- 1782 160. B. Sahiner, H.-P. Chan, N. Petrick, R.F. Wagner, and L. Hadjiiski, "Feature selection and  
1783 classifier performance in computer-aided diagnosis: The effect of finite sample size,"  
1784 *Medical Physics*, 27, 1509-1522 (2000).
- 1785 161. X. Bouthillier, C. Laurent, and P. Vincent, "Unreproducible research is reproducible," in  
1786 *36th International Conference on Machine Learning, ICML 2019*, pp. 1150-1159 (2019).
- 1787 162. M. McDermott, S. Wang, Marinsek, N. Ranganath, R. Ghassemi, and L. M. Foschini,  
1788 "Reproducibility in machine learning for health," arXiv preprint arXiv:1907.01463,  
1789 (2019).
- 1790 163. S.N. Goodman, D. Fanelli, and J.P.A. Ioannidis, "What does research reproducibility  
1791 mean?," *Science Translational Medicine*, 8, 341ps12 (2016).
- 1792 164. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model  
1793 uncertainty in deep learning," in *33rd International Conference on Machine Learning,*  
1794 *ICML 2016*, pp. 1651-1660 (2016).
- 1795 165. A. Kendall and Y. Gal, *What Uncertainties Do We Need in Bayesian Deep Learning for*  
1796 *Computer Vision?*, in *Advances in Neural Information Processing Systems 30*, I. Guyon,  
1797 *et al.*, Editors. 2017,
- 1798 166. R. Robinson, V.V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M.M. Sanghvi, N.  
1799 Aung, J.M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A.M. Lee, V. Carapella, Y.J. Kim,  
1800 S.K. Piechnik, S. Neubauer, S.E. Petersen, C. Page, P.M. Matthews, D. Rueckert, and B.  
1801 Glocker, "Automated quality control in image segmentation: application to the UK

- 1802 Biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular*  
1803 *Magnetic Resonance*, 21, (2019).
- 1804 167. Y. Yang, X. Guo, Y. Pan, P. Shi, H. Lv, and T. Ma, "Uncertainty Quantification in  
1805 *Medical Image Segmentation with Multi-decoder U-Net*," arXiv preprint  
1806 arXiv:2109.07045, (2021).
- 1807 168. M. Rezaei, J. Näppi, B. Bischl, and H. Yoshida, "Bayesian uncertainty estimation for  
1808 *detection of long-tail and unseen conditions in abdominal images*," *Proc of SPIE Medical*  
1809 *Imaging*, 12033, 1203311 (2022).
- 1810 169. Z. Salahuddin, H.C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep  
1811 *neural networks for medical image analysis: A review of interpretability methods*,"  
1812 *Computers in Biology and Medicine*, 140, (2022).
- 1813 170. M. Reyes, R. Meier, S. Pereira, C.A. Silva, F.-M. Dahlweid, H. von Tengg-Kobligk, R.M.  
1814 *Summers*, and R. Wiest, "On the Interpretability of Artificial Intelligence in Radiology:  
1815 *Challenges and Opportunities*," *Radiology. Artificial intelligence*, 2, e190043-e190043  
1816 (2020).
- 1817 171. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Mueller, "Evaluating the  
1818 *Visualization of What a Deep Neural Network Has Learned*," *Ieee Transactions on Neural*  
1819 *Networks and Learning Systems*, 28, 2660-2673 (2017).
- 1820 172. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, *Learning Deep Features for*  
1821 *Discriminative Localization*, in *IEEE Conference on Computer Vision and Pattern*  
1822 *Recognition (CVPR 2016)*. 2016, p. 2921-2929.
- 1823 173. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, *Grad-CAM:*  
1824 *Visual Explanations from Deep Networks via Gradient-based Localization*, in *2017 Ieee*  
1825 *International Conference on Computer Vision*. 2017, p. 618-626.
- 1826 174. H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-  
1827 *CAM: Score-weighted visual explanations for convolutional neural networks*," in *IEEE*  
1828 *Computer Society Conference on Computer Vision and Pattern Recognition Workshops*,  
1829 pp. 111-119 (2020).
- 1830 175. A.J. Barnett, F.R. Schwartz, C. Tao, C. Chen, Y. Ren, J.Y. Lo, and C. Rudin, "A case-  
1831 *based interpretable deep learning model for classification of mass lesions in digital*  
1832 *mammography*," *Nature Machine Intelligence*, 3, 1061-+ (2021).
- 1833 176. N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J.  
1834 *Patel*, M. Gidwani, J. Adebayo, M.D. Li, and J. Kalpathy-Cramer, "Assessing the  
1835 *Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging*,"  
1836 *Radiology. Artificial intelligence*, 3, e200267-e200267 (2021).
- 1837 177. M.T. Ribeiro, S. Singh, C. Guestrin, and M. Assoc Comp, "*Why Should I Trust You?*"  
1838 *Explaining the Predictions of Any Classifier*. *Kdd'16: Proceedings of the 22nd Acm*  
1839 *Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016. 1135-  
1840 1144.
- 1841 178. H.P. Chan, R.K. Samala, L.M. Hadjiiski, and C. Zhou, *Deep Learning in Medical Image*  
1842 *Analysis*, in *Deep Learning in Medical Image Analysis: Challenges and Applications*, G.  
1843 *Lee and H. Fujita*, Editors. 2020, p. 3-21.
- 1844 179. H.-P. Chan, L.M. Hadjiiski, and R.K. Samala, "Computer-aided diagnosis in the era of  
1845 *deep learning*," *Medical Physics*, 47, e218-e227 (2020).
- 1846 180. T.W. Freer and M.J. Ulissey, "Screening mammography with computer-aided detection:  
1847 *Prospective study of 12,860 patients in a community breast center*," *Radiology*, 220, 781-  
1848 786 (2001).

- 1849 181. M.A. Helvie, L. Hadjiiski, E. Makariou, H.-P. Chan, N. Petrick, B. Sahiner, S.C.B. Lo, M.  
1850 Freedman, D. Adler, J. Bailey, C. Blane, D. Hoff, K. Hunt, L. Joynt, K. Klein, C.  
1851 Paramagul, S.K. Patterson, and M.A. Roubidoux, "Sensitivity of noncommercial  
1852 computer-aided detection system for mammographic breast cancer detection: Pilot clinical  
1853 trial," *Radiology*, 231, 208-214 (2004).
- 1854 182. R.L. Birdwell, P. Bandodkar, and D.M. Ikeda, "Computer-aided detection with screening  
1855 mammography in a university hospital setting," *Radiology*, 236, 451-457 (2005).
- 1856 183. J.C. Dean and C.C. Ilvento, "Improved cancer detection using computer-aided detection  
1857 with diagnostic and screening mammography: Prospective study of 104 cancers,"  
1858 *American Journal of Roentgenology*, 187, 20-28 (2006).
- 1859 184. M.J. Morton, D.H. Whaley, K.R. Brandt, and K.K. Amrami, "Screening mammograms:  
1860 Interpretation with computer-aided detection - Prospective evaluation," *Radiology*, 239,  
1861 375-383 (2006).
- 1862 185. F.J. Gilbert, S.M. Astley, M.G. Gillan, O.F. Agbaje, M.G. Wallis, J. James, C.R. Boggis,  
1863 S.W. Duffy, and C.I. Grp, "CADET II: A prospective trial of computer-aided detection  
1864 (CAD) in the UK Breast Screening Programme," *Journal of Clinical Oncology*, 26, 508  
1865 (2008).
- 1866 186. D. Regge, P. Della Monica, G. Galatola, C. Laudi, A. Zambon, L. Correale, R. Asnagli,  
1867 B. Barbaro, C. Borghi, D. Campanella, M.C. Cassinis, R. Ferrari, A. Ferraris, R. Golfieri,  
1868 C. Hassan, F. Iafrate, G. Iussich, A. Laghi, R. Massara, E. Neri, L. Sali, S. Venturini, and  
1869 G. Gandini, "Efficacy of Computer-aided Detection as a Second Reader for 6-9-mm  
1870 Lesions at CT Colonography: Multicenter Prospective Trial," *Radiology*, 266, 168-176  
1871 (2013).
- 1872 187. J. Concato, N. Shah, and R.I. Horwitz, "Randomized, controlled trials, observational  
1873 studies, and the hierarchy of research designs," *New England Journal of Medicine*, 342,  
1874 1887-1892 (2000).
- 1875 188. D. Gur, J.H. Sumkin, H.E. Rockette, M. Ganott, C. Hakim, L. Hardesty, W.R. Poller, R.  
1876 Shah, and L. Wallace, "Changes in breast cancer detection and mammography recall rates  
1877 after the introduction of a computer-aided detection system," *Journal of the National  
1878 Cancer Institute*, 96, 185-190 (2004).
- 1879 189. J.J. Fenton, L. Abraham, S.H. Taplin, B.M. Geller, P.A. Carney, C. D'Orsi, J.G. Elmore,  
1880 W.E. Barlow, and C. Breast Canc Surveillance, "Effectiveness of Computer-Aided  
1881 Detection in Community Mammography Practice," *Journal of the National Cancer  
1882 Institute*, 103, 1152-1161 (2011).
- 1883 190. M. Gromet, "Comparison of computer-aided detection to double reading of screening  
1884 mammograms: Review of 231,221 mammograms," *American Journal of Roentgenology*,  
1885 190, 854-859 (2008).
- 1886 191. C.D. Lehman, R.D. Wellman, D.S.M. Buist, K. Kerlikowske, A.N.A. Tosteson, D.L.  
1887 Miglioretti, and S. Breast Canc, "Diagnostic Accuracy of Digital Screening  
1888 Mammography With and Without Computer-Aided Detection," *Jama Internal Medicine*,  
1889 175, 1828-1837 (2015).
- 1890 192. S. Cruz Rivera, X. Liu, A.-W. Chan, A.K. Denniston, M.J. Calvert, A.I. Spirit, C.-A.W.  
1891 Group, A.I. Spirit, C.-A.S. Group, A.I. Spirit, and C.-A.C. Group, "Guidelines for clinical  
1892 trial protocols for interventions involving artificial intelligence: the SPIRIT-AI  
1893 extension," *Nature medicine*, 26, 1351-1363 (2020).
- 1894 193. X. Liu, S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, A.I. Spirit, and C.-A.W.  
1895 Group, "Reporting guidelines for clinical trial reports for interventions involving artificial  
1896 intelligence: the CONSORT-AI extension," *Nature medicine*, 26, 1364-1374 (2020).

1897