AI and Ethics
https://doi.org/10.1007/s43681-022-00161-9

ORIGINAL RESEARCH

Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability

Alvaro Fernandez-Quilez[1,2,3]

Received: 9 January 2022 / Accepted: 9 April 2022

Abstract

AI systems are quickly being adopted in radiology and, in general, in healthcare. A myriad of systems is being proposed and developed on a daily basis for high-stake decisions that can lead to unwelcome and negative consequences. AI systems trained under the supervised learning paradigm greatly depend on the quality and amount of data used to develop them. Nevertheless, barriers in data collection and sharing limit the data accessibility and potential ethical challenges might arise due to them leading, for instance, to systems that do not offer equity in their decisions and discriminate against certain patient populations or that are vulnerable to appropriation of intellectual property, among others. This paper provides an overview of some of the ethical issues both researchers and end-users might meet during data collection and development of AI systems, as well an introduction to the current state of transparency, interpretability and explainability of the systems in radiology applications. Furthermore, we aim to provide a comprehensive summary of currently open questions and identify key issues during the development and deployment of AI systems in healthcare, with a particular focus on the radiology area.

Keywords Deep learning · Healthcare · Radiology · Data ethics · Algorithms

✉ Alvaro Fernandez-Quilez
alvaro.f.quilez@uis.no

[1] Department of Quality and Health Technology, University of Stavanger, Stavanger, Norway

[2] Stavanger Medical Imaging Laboratory, Department of Radiology, Stavanger University Hospital, Stavanger, Norway

[3] Centre for Age-Related Medicine (SESAM), Stavanger University Hospital, Stavanger, Norway

Published online: 28 April 2022

1 Introduction

The convergence of medical imaging with artificial intelligence (AI) is rapidly enabling the development of tools that are able to automatize tasks that have traditionally been carried out by human experts [1]. In particular, deep learning (DL) techniques have led to important breakthroughs in the computer vision area [2], which have been successfully applied to the radiology domain to, for example, classify patients based on chest X-rays diagnosis [3]—a key clinical focus area [4]—or nodule detection in computed tomography (CT) images [5], enabling the field to develop tools in times of crisis (COVID19) in a timely manner [6]. The DL field has quickly progressed during the past years, and as the field keeps evolving, medical imaging systems powered by DL algorithms will become more widespread leading to a larger number of radiology data analysis results that will depend on the performance and efficiency of such systems [7].

The availability and accessibility to large amounts of human-annotated data has been one of the key elements driving the quick growth and success of traditional DL-systems [4, 8]. Notably, there are big challenges associated to radiology data collection, annotation process, availability and accessibility which lead to data scarcity and ultimately, hinders the ability of the DL-powered tools to support the radiologist in their decisions [9]. Furthermore, if not properly assessed and reflected upon, the data collection challenges and limitations can cause other potential issues such as data bias, promoting or harming unrepresented groups based on their gender, sexual orientation, ethnic, social or economic factors, among others [10–12]. In particular, the radiology field is also specially hindered by technical acquisition factors bias, due to differences between different machines or acquisition techniques [9]. As data bias is recognized as a force to reckon with, DL explainability techniques have gained popularity with the promise of ensure safety, efficacy and equity in the deployment of DL-powered systems [12, 13]. However, it has been argued that in their current form, explainable techniques for DL applications might not be suitable for scenarios such as the radiology one in which patients' lives might be at stake [13].

The aim of this work is to provide a non-exhaustive overview of the challenges associated with data collection and usage in radiology DL-systems as well as to promote and increase awareness of the issues that might arise in the deployment phase of the developed DL-system if the impact of all the steps involved in the collection process and data usage are not properly assessed nor reflected upon. In addition, the work provides a critical view of the current explainable DL techniques and their added value to avoid the opacity of DL algorithms avoiding the so-called "black box" nature.

2 Data collection in radiology research

Universally lamented by researchers as an overlooked aspect of research, data collection is, arguably, one of the most important factors for the success of DL-systems in a clinical environment [14, 15]. Researchers usually do not pay enough attention to the data collection details and critical aspects such as a good study design, how the data will be collected, data availability, design of the data collection system and establishing a quality control are often neglected or ignored [16, 17]. As the quality and amount of data play a crucial role in the results obtained by DL-powered radiology systems trained with the supervised learning paradigm approach—the most common learning paradigm in radiology DL [18]—, taking into account those aspects are of special relevance to avoid undesired outcomes such as results that might reflect biases present in the data.

Avoiding all the problems that could arise during the data collection process—assuming a previously defined good study design—represents a challenge but it is the researchers' responsibility to try to limit the reasons for poor data quality or biased data and their potentially negative impact on the DL-system. To ensure representative and high-quality data, continuous monitoring of the data is required during the data extraction process [14, 17]. In current practice, information system providers (e.g., PACS) have attempted to automatize data extraction providing a unified system to store, transmit and retrieve data with the main objective of optimizing the radiologist workflow. The reality is that the lack of standardized data and contextual and user-specific variability might hinder the extraction ability of the information system [19]. Hereby, long delays in data retrieval or unexpected difficulties should be taken into consideration when planning the data collection.

To proceed with the data collection, local institutional review board (IRB) approval is a must in most, if not all studies involving clinical data. Irrespectively of the nature of the study, data collection should only start after all ethical and legal procedures are in place [17]. If the intention is to deploy the DL-system in clinical practice, the CE mark is a pre-requisite for medical devices to be allowed in the market in Europe. Furthermore, clearance by the U.S. Food & drugs administration (FDA) might be required [18]. A detailed and comprehensive list of the current CE and FDA cleared products and more details about CE and FDA marks can be found in [20, 21]. Among the ethical, legal and procedural aspects that need to be addressed before getting IRB, CE and FDA approval, we can find patient privacy, data protection, informed consent, data bias, data "truthfulness", data ownership, providing meaningful and moral access to data and DL-system transparency, interpretability and explainability [14, 22]. One could argue that the amount of control of such a large number of regulatory agencies could prevent the use of DL in radiology and have a negative effect on the integration and smooth adaptation of new DL-powered technologies in radiology. However, in the current state of DL-systems development and the lack of procedures that provide and ensure a safe deployment (inclusive technologies, for instance as introduced in the following sections) of DL-technologies that have been approved by international agencies [21, 22] the breach between the theoretical demands of those systems and the reality during their deployment and potential issues in a clinical use seems too great to relax the existing control. On the other hand, new measures might be required to meet the needs of clinical practice instead of needs coming from a developer and more technological point of view.

3 Ethical issues related to data in radiology

Inter-disciplinary teams including radiologists working on or with DL-powered systems have a moral duty to use the data they collect in such a way that is for a common good and to improve radiology practices. Likewise, they have a duty to not use data in a way that may have a negative impact or discriminate the patients [14]. Due to the large dependence of DL-systems on data amount and quality—well-labeled and high-quality "ground truths", the data used to define the quality of the results during the training of the algorithm and after the training process—are highly sought after and its value is skyrocketing. Besides the common good that is achievable by making a good and ethical use of the available data, harm can also be caused if unethical use of data nor following IRB, CE or FDA requirements is not in place.

Unethical use of the data can be unintentional due to challenges associated to its collection, such as difficulties to obtain patient consent. Limitations during data collection can have a negative effect on the quality and diversity of the collected data, resulting in data biases the researcher might not be aware of, with serious consequences such as the discrimination of small groups by the developed algorithm [12, 23].

3.1 Data bias and shift

Data bias happens to some degree in any collected data [14], and it can be defined as the differences in performance of the algorithm when dealing with subpopulations of different characteristics (e.g., ethnical, economical or technical). Even though DL algorithms have been shown to potentially reduce bias and improve healthcare practices or workflows [24], application of DL algorithms has also been shown to systemize or amplify biases [11, 25].

Of particular interest and prevalence is the selection or sampling bias, which occurs when the data collection does not represent the population accurately [26]. Selection or sampling bias is usually the result of aforementioned limitations during data collection, with the result of using data that is available at the time the DL-system is planned to be developed. A really common example of selection bias is when data coming from a single institution is used to develop and train the DL-algorithm resulting in a discrimination of underrepresented subsets of other institutions' populations [27], which can result in the deployment of a system that underperforms and discriminates population of underrepresented characteristics in the original institution dataset when adopted in an institution or setting with a different acquisition protocol or data characteristics. Such a situation is a recurring issue in developed DL-systems, as even when clear evaluation and validation methodologies are in place for similar areas such as classic statistical methods, DL-powered systems suffer from the lack of validation protocols globally accepted and implemented by researchers and developers [25–27] such as external validation. A clear example of this existing issues is the aforementioned external validation, which uses representative data of other institutions to avoid issues such as selection bias in future deployments and ensure the generalization of the results, and in spite of its relevance, only 6% of the recent medical DL-papers included validation on an independent external data [28].

Data shift is a subset of selection bias and poses one of the biggest threats to generalization and usage of DL-systems. Data shift usually occurs because the data used to train the DL-system (commonly of retrospective nature), does not accurately reflect the characteristics of the data that will be used with the developed system in the future. While for a radiologist is common to assess and take into account technical differences in the acquisition of the data such as slice thickness or scanner brand, the developed DL-systems lack the ability to detect those differences if they have not been taken into account in the training phase of the model [14]. To a certain extent, one could say that radiologists are able to re-train themselves to adapt to data shifts while a DL-system requires careful assessment of the steps followed to train it if planned to be developed in an environment in which data shifts might be present. In such a scenario, re-assessment of the system should be performed as the original results might not accurately reflect the system's performance in the presence of data shifts [14, 25]. A potential solution is data harmonization a technical process that allows to unify and align data of different sources and characteristics. Data harmonization can be accomplished by setting general rules and guidelines which allows such a unification of the data characteristics across institutions and vendors, which would allow for a safe deployment of models trained with a single (but generic technical characteristics) institution data and would allow for easier multi-institutional model developments. Despite its usefulness, data harmonization would require complementary practices and, for instance, a thorough analysis of the data population characteristics would need to be in place. Nevertheless, in its current state, data harmonization lacks general guidelines or policies which could benefit the process and the avoidance of data shifts with its implementation.

Data biases and shifts can be especially harmful when trying to deploy DL-powered models that have been trained on data from developed countries and that might not be representative of rural areas, limiting the usefulness and transferability of the models in areas that could greatly benefit from them due to the lack in resources, specialists and data. This kind of setting leads to a situation in which DL-models are only accessible to those that have enough resources (in the form of technology, specialists and, overall, and wealthiness) and access to large amounts of data while limiting their usability and amplifying biases present in the original institution (the one that has developed the model) in, for instance, a rural area due to the presence of data bias, shifts and validation of the system in scenarios in which arguably, the need for those systems is greater than in areas that enjoy that resource availability, presenting an important moral and ethical dilemma. Situations such as the presented one raises several concerns of whether DL-powered systems are double-edged swords, with a great potential to improve the way healthcare is delivered but at the same time with the ability to worsen disparities in healthcare. Evidence found in studies shows that this is not just a theoretical concern. Studies have found different underdiagnosis rates depending on the race, ethnicity, sex, age and insurance type [10–12]. Specifically, factors such as insurance can be highly influential in certain contexts, such as the US where greater disparities exist among the population [29] while not as relevant for other areas. However, as AI systems become more widely available, it is to be expected some degree of transferability and re-usability of the systems developed in the leading countries in AI development [30], which can lead to greater disparities and amplification of the existing ones in countries or regions where those disparities were not present. In other countries and cases where insurance is not as relevant, factors such the ethnicity might be prone to creep into the system and affect the performance of the system in the presence of underrepresented groups due to inherent differences in the structures of interest present in the image [31]. Other studies have found significant differences in the detection ability of algorithms depending on the skin tone and gender of the population [28], which makes healthcare providers wonder what would happen if that kind of algorithms were, for instance, to be used to diagnose melanoma on light versus dark skin [32].

In radiology, and generally speaking in healthcare, data that are left unchecked and not properly analyzed could incorporate and perpetuate economic and social biases that already contribute to healthcare disparities, especially in situations and conditions with complex trade-offs and outcomes with high uncertainty. For instance, if patients with low income tend to do worse when receiving treatment, DL-systems could recommend against treating them, as they learn from the data characteristics used to trained the system. Hence, social biases that contribute to healthcare disparities that are left without rigorous analysis in the implementation phase of the algorithm can contribute to decisions and systems that perpetrate them, such as the example presented above. Furthermore, they can be amplified by extrapolating those decisions to other institutions or settings in which the system might be deployed DL-systems pose the risk to automatize and make biases invisible that are otherwise well known if rigorous analysis of data used to train the system is not in place, and the DL-system decisions are accepted over our moral and knowledge-guided intuition.

3.2 Data ownership, recollection and model re-training

As data sharing practices start to become more common and prevalent among radiologists and other healthcare practitioners (among others), data ownership questions need to be addressed. "Who owns the data?" is a question that has already been addressed by regulatory bodies with the result of a different type of answer depending on the country the question is formulated [14]. While there is no general consensus among different hospitals, many include a clause in their general consent form given to the patient, which allows to use the data in a retrospective way for research purposes and which is generally accepted by patients [33]. In Europe, the General Data Protection Regulation (GDPR) states that patients own and control their sensitive, personal or identifiable data (medical and non-medical). The GDPR allows the patient to withdraw consent to use their data at any time and requires to obtain consent every time the data is planned to be reused or shared [34]. These legal restrictions are different in other parts of the world, such as the United States. Such a difference between the legal restrictions might limit the development of DL-systems in some parts of the world, skewing the availability and equal opportunities for all the entities involved in the field.

Practices such as DL-system re-training also illustrate the need for discussions on data ownership and re-use. For instance, should the patients that gave consent to use their data and train a DL-system be notified if that system is reused with the parameters obtained from that training in another institution and re-trained with other data? In the same scenario, what happens if a patient decides to withdraw their consent? Should the model still be used with the parameters obtained trained using the patients' data or should it be re-trained without it? Data ownership is loosely defined in some scenarios that have raised with the disruption of DL-systems and requires further discussion and attention.

3.3 Data privacy and sharing

The increasing demand and growth of developed DL-systems in radiology are pushing the limits of data availability and sharing. The unprecedented value of medical data to build DL-powered systems is blurring the line between research-only and commercial use of it, raising some concerns around the regulations and policies involved in academic and commercial data usage [14]. If a company buys the rights to medical data access and makes profit out of a product built using it, who should benefit from the profits? Since patients have the right and retain in the majority of the cases access to their data, should they also be accounted for during the profit sharing? Can they, for instance, refuse to agree on selling their data to the company but allow to using it for academic use? Could they refuse to sell to the specific company but sell it to a different one? The lack of mechanisms and rules make it hard to answer those questions and there is a growing need for consensus and assessment of potential situations that right now are vaguely defined from the point of view of the data ownership [35]. Hence, there is a need for an updated general guideline that contemplates situations such as the aforementioned ones.

Anonymization of the data is an important step to ensure patients' privacy rights are preserved as well as to allow for data sharing practices. Nevertheless, full anonymization of the data is more complicated than one could picture at first. As technologies such as AI keep improving, re-identification of the data is becoming easier, such that the source of the data is easily identifiable. For example, it has been found that facial recognition to 3D reconstruction can recover the original identification of the data [36, 37] by generating realistic facial reconstructions from deidentified medical images (e.g., MRI) and identifying the participants by matching them to publicly available photographs of named persons (e.g., from social networks). Given the unprecedented ability of DL techniques to re-identify data from features that would apparently look harmless in the eyes of other techniques or from a human perspective, additional developments in privacy-enhancing techniques are required to avoid such an identification ability. Specific software to avoid re-identification of the data has been developed with success [36], but the solutions are tailored to specific models and data (e.g., CT and specific DL-models based on convolutional neural networks), thus limiting their generalization to other potential re-identification systems. Another cause for concern related to data sharing challenges is the ability of large data companies that control both social medial and medical AI systems, such that they are able to gather data from daily-use tools such as smartphones and match it with social media one, which could, arguable, be considered a non-moral use of the data. The ability to monetize data is leading to a model of self-governance by those who own the data and actions that should not be morally acceptable. Generally speaking, most of the daily users of social media platforms are unaware and undervalue the monetary value of the data they provide to such large companies or healthcare providers. Hypothetically speaking, a necessary condition should be, therefore, an informed consent by the end-user or patient only when awareness has been raised of the economic value carried by their data. However, if users where to give consent for all our online activity an unreasonable amount of time would be spent reading terms and conditions. Hence, new approaches that allow for an easy integration in different platforms and acceptance of the implications that data sharing has for the end-user would be desirable.

Freely accessible radiology data could benefit for the greater good of patients and society, and new necessities in the form of robust infrastructures to share radiological data in a safe and preserving patients' privacy rights are driving new technological developments aiming to solve those challenges [14, 38]. Two of the most iconic developments are systems that use the federated learning (FL) paradigm and blockchain models. Federated learning is a learning paradigm that allows to train models "in-house" while still allowing in an indirect way to update the model parameters based on other centers' data. The key element in a federated learning setting is that instead of sharing the data, the users share the parameters of the model and update their own "in-house" model accordingly, avoiding privacy concerns associated to patients' data. In a similar fashion, blockchain technologies promise to deliver a secure methodology and encryption such that access to medical data is secure across different sites [14]. Nevertheless, in the case of blockchain technologies, the state of play is still immature. A large number of recent research papers present novel blockchain frameworks, architectures or models from a conceptual point of view but there is rarely an pilot implementation or prototype showing promising results in a real-world scenario [39]. In addition, the deployment of block chain technology in health at a big scale (national, for instance) is scarce. Some of those examples include the deployment of the technology in countries such as Estonia and Malta with promising applications in identity management and patient consent, among others [39]. In spite of the promising perspectives blockchain offers, more development and research are required to assess the positive impact of it in clinical practice. Regarding FL, like any other AI-based model they are prone to attacks (otherwise called "hacking") which can result in disruption of the services or compromised data [40]. In particular, the attacks can be introduced by a compromised central server or a compromised local device in the learning framework or by any participant in the FL workflow [40]. In FL, attacks can be particularly critical when compared to other settings due to the distributed nature of the technology, as it makes it harder to deploy defense measures in the event of attacks [40]. Hence, more research towards the defense against potential attacks is required before the technology is fully ready to be deployed and adapted without the risk of compromising highly sensitive data.

3.4 DL-generated data (synthetic data)

DL architectures such as Generative Adversarial Networks (GAN) [41] have gained a considerable amount of attention due to their ability to generate synthetic data that resembles the one used to train the architecture after learning the distribution (characteristics) of the original data. The motivation for the usage of synthetic data comes from the currently limited availability of radiology datasets, along with the data collection and sharing challenges [14]. As an example of synthetic-generation data, one could generate realistic-looking prostate magnetic resonance images (MRI) after training the GAN with thousands of prostate MRI slices [42]. Following, the generated images can be used to train the DL-system along with the original data, considerably increasing the amount of available data to train with (hopefully) a positive effect on the final performance of the developed model.

After the model able to generate synthetic data has been trained, generating new data become fast and inexpensive. Synthetic data can be particularly useful for pre-training [8], trying to tackle potential biases such as data imbalance (for instance, a clear higher prevalence of healthy population when compared with the amount that has a specific diagnostic in the data collection) by increasing the amount of data of the least prevalent class or ground truth [43] or for new learning paradigms such as self-supervised learning [44]. In addition, synthetic data minimize the risk of compromising patient data and the challenges associated with data sharing practices. By contrast, synthetic data can also introduce artifacts which are impossible to perceive by the human eye, hindering the final performance of the model. Furthermore, if the original data used to develop the generative model suffer from biases, the synthesized data will pick up and amplify those data biases. Little research has been carried out to understand the effect of synthesized images on real-life settings, including the current evaluation measures to determine whether the synthetic data reach an acceptable quality standard. For instance, evaluation measures commonly used for natural images [45] might not be enough to evaluate synthetic data generated with the purpose of screening patients for a certain disease and other measures to evaluate specific areas or regions of, for instance, the generated image (e.g., a tumor), would be of interest.

4 DL-systems' transparency, interpretability and explainability

Transparency, interpretability and explainability are concepts that are closely related to data ethics and necessary to build trust in DL-system and in its safe usage, and for the patients' and society benefit. It is in human nature to have the need to understand how decisions are made and which are the factors underlying those decisions. More importantly, if a DL-system contributes to adverse events, the team involved in its development needs to be able to understand and pinpoint why the system reached such a result and how it reached it, such that reasonable explanations can be given to the ones affected by the adverse outcome [14]. Moreover, by per GDPR regulations, an individual has the right to an explanation of how the system reached that decision if an automated decision-making system is being used [34, 46]. However, the extent of such an "explanation" is unclear as the European Council Data Protection Working Party defines it as "the right to the envisaged consequences of a process", rather than an explanation of a particular decision [14, 47].

The concept of "black box" for DL-systems has been present ever since their disruption and historically, DL-systems have lacked mechanisms that allowed to understand why they obtain certain results. The "black box" nature can result especially problematic in healthcare and in particular, radiology. Take for instance a DL-system that recommends a diagnosis to patients based on their imaging data. Would a patient undergo treatment or surgery if the doctor mentioned "my computer software has proposed a specific diagnosis that would require further treatment, so you should take it" without any other reasonable explanation that led to that decision? A human radiologist will usually be able to explain the train of thoughts behind a decision [48]. Similarly, we require mechanisms that to some degree allow to have some traceability and explainability for DL-systems decisions [49, 50]. Some initiatives in explainable AI (XAI) are already in place. For example, saliency maps have been proposed to highlight the areas of the image deemed most important, for instance, for the diagnosis of a certain disease [51]. Nevertheless, as noted in [13], such highlights contain both useful and non-useful information (from the human perspective) and the region does not reveal exactly what it was in that area the model considered useful for the diagnosis. At this point, it is important to remember that a DL-system is not human but rather something built by humans. Hereby, asking for the same level of traceability a human could provide should be out of the scope of the capabilities of a DL-system. However, a DL-system should be capable of providing enough information such that a radiologist or health specialist equipped with the necessary knowledge is able to interpret and provide a human-like explanation of the decisions reached by the computer.

Interpretability can be defined as the ability to understand the workings of an AI model, while transparency occurs when the model is both visible and comprehensible to outside viewers [14]. In spite of being desirable attributes for the developed DL-systems, the more transparent and interpretable a model is, the more prone it might be to appropriation of intellectual property or malicious attacks [35, 52]. Generally speaking, the more complex a system is, the less transparent and interpretable it might become, due to the large number of parameters and operations happening under the hood of the developed systems. Hereby, finding the right balance between model complexity, interpretability and transparency is especially important, such that the final performance of the model is not hindered nor privacy is compromised. Guidelines from the radiology community might be needed to explain and assess DL-systems with specifications on the amount of knowledge required by the radiologist using the DL-system and what adequate explanations entail.

5 Conclusions

This paper has shown that it is important to evaluate how the data collection process influences the final performance of the DL-system and the data ethics associated to it. We present an introduction to the data collection process challenges present in the radiology area that can hinder the amount and quality of the data available for DL-systems development. Following, some data ethics points that are deemed of interest and of relevance in the context of data collection are presented. In particular, we focus on data bias and shifts, ownership, recollection, privacy, sharing, synthetic data and model re-training. Ethical problems can arise from the challenges experienced during data collection and which can have undesirable outcomes such as the underdiagnosis of subpopulations based on their ethnical origin or economic status, potentially exacerbating the disparities that are already present in healthcare services. We argue that is the duty of the DL-system developers to take into account these potential issues such that the final result could benefit for the greater good the patients, instead of perpetrating issues that are already present in nowadays' practices.

The widespread of DL-systems in radiology also calls for new approaches to provide some degree of interpretability, transparency and explainability. Current efforts of explainability in AI are not necessarily aligned with the necessities in healthcare [13] and the trade-off between model complexity and transparency/interpretability should be carefully assessed, as data confidentiality and model performance might be compromised if not properly addressed. Questions such as "what is an appropriate explanation in healthcare?", "To what extent should practitioners be able to dissect and understand the model?" need more discussion and areas such as radiology might require guidelines to provide some general and accepted consensus among practitioners and users of DL-systems.

Artificial intelligence, and in particular, DL-systems, hold tremendous potential to improve radiology workflow and, in general, medicine. If properly deployed and developed, more efficient, accurate and equitable outcomes could be obtained. Nevertheless, such potential requires awareness of the ethical issues that might arise during its development and deployment, as well as anticipation and guarding against potential negative consequences of it. Ultimately, humans are the ones responsible of the design and development of DL-systems and at the same time, responsible for the patient care. Hereby, it is our duty to ensure an ethical and for the general good usage of AI tools.

Funding Open access funding provided by University Of Stavanger. Universitetet i Stavanger.

Declarations

Conflict of interest The corresponding author states that there is no conflict of interest.

References

1. Topol, E.: Deep medicine: how artificial intelligence can make healthcare human again, 1st edn. Basic Books Inc, New York (2019)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2017). https://doi.org/10.1145/3065386
3. Irvin J, et al. (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In Proc. of the AAAI Conference on Artificial Intelligence. 33:590–597.
4. Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R.: Deep learning-enabled medical computer vision. NPJ Digit. Med. **4**(1), 1–9 (2021)

5. Ding, J., Li, A., Hu, Z., Wang, L.: In medical image computing and computer assisted intervention—MICCAI, pp. 559–567. Springer, Cham (2017)
6. Zhang, J., Xie, Y., Pang, G., Liao, Z., Verjans, J., Li, W., Sun, Z., He, J., Li, Y., Shen, C., Xia, Y.: Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. IEEE Trans. Med. Imaging **40**(3), 879–890 (2020)
7. Jarrah, M.H.: Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. Bus. Horiz. **61**(4), 577–586 (2018)
8. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. arXiv preprint arXiv:1902.07208. 2019 Feb 14.
9. WHOQoL Group: Study protocol for the World Health Organization project to develop a quality of life assessment instrument (WHOQOL). Qual. Life Res. **2**, 153–159 (1993)
10. Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat. Med. **10**, 1–7 (2021)
11. Larrazabal, A.J., et al.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl Acad. Sci. USA **117**, 12592–12594 (2020)
12. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. InBIOCOMPUTING 2021: Proceedings of the Pacific Symposium (2020) pp. 232–243.
13. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. **3**(11), e745–e750 (2021)
14. Geis, J.R., Brady, A.P., Wu, C.C., Spencer, J., Ranschaert, E., Jaremko, J.L., Langer, S.G., Kitts, A.B., Birch, J., Shields, W.F., van den HovenGenderen, R.: Ethics of artificial intelligence in radiology: summary of the joint European and north American multisociety statement. Can Assoc Radiol J. **70**(4), 329–334 (2019)
15. Crewson, P.E., Applegate, K.E.: Data collection in radiology research. Am. J. Roentgenol. **177**(4), 755–761 (2001)
16. Friedman LM, Furberg CD, Demets DL (1998) Data collection and quality control in: fundamentals of clinical trials. Springer, New York.
17. Altman, D.G.: Statistics and ethics in medical research: collecting and screening data. BMJ **281**, 1399–1401 (1980)
18. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. **1**(54), 280–296 (2019)
19. Reiner, B.: Strategies for medical data extraction and presentation part 1: current limitations and deficiencies. J. Digit. Imaging **28**(2), 123–126 (2015)
20. van Leeuwen, K.G., Schalekamp, S., Rutten, M.J., van Ginneken, B., de Rooij, M.: Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur. Radiol. **31**(6), 3797–3804 (2021)
21. Benjamens, S., Dhunnoo, P., Meskó, B.: The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. **3**(1), 1–8 (2020)
22. Mittelstadt, B.D., Floridi, L.: The ethics of big data: current and foreseeable issues in biomedical contexts. Sci Eng Ethics **22**(2), 303–341 (2016)
23. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
24. Chen, I.Y., Joshi, S., Ghassemi, M.: Treating health disparities with artificial intelligence. Nat. Med. **26**, 16–17 (2020)
25. Wiens, J., et al.: Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. **25**, 1337–1340 (2019)
26. Group SI, Community, F.R.: Artificial intelligence and medical imaging 2018: French radiology community white paper. Diagn. Interv. Imaging **99**(11), 727–742 (2018)
27. Kim, D.W., Jang, H.Y., Kim, K.W., et al.: Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol **20**, 405–410 (2019)
28. Hardesty, L.: Study finds gender and skin-type bias in commercial artificial-intelligence systems. Retriev April. **11**(3), 2019 (2018)
29. Chen RJ, Chen TY, Lipkova J, Wang JJ, Williamson DF, Lu MY, Sahai S, Mahmood F. Algorithm fairness in AI for medicine and healthcare. arXiv preprint arXiv:2110.00603. (2021 Oct 1).
30. Savage, N.: The race to the top among the world's leaders in artificial intelligence. Nature **588**(7837), 102–104 (2020)
31. Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning-based segmentation. In international conference on medical image computing and computer-assisted intervention, pp. 413–423. Springer, Cham (2021)
32. Khullar, D.: Opinion A.I. could worsen health disparities. N.Y. Times, New York (2019)
33. Wendler, D.: One-time general consent for research on biological samples: is it compatible with the health insurance portability and accountability act? Arch. Intern. Med. **166**(14), 1449–1452 (2006)
34. Council of Europe. Convention for the Protection of individuals with regard to Automatic Processing of Personal Data. (1985).
35. Brady, A.P., Neri, E.: Artificial intelligence in radiology—ethical considerations. Diagnostics. **10**(4), 231 (2020)
36. Mazura, J.C., Juluru, K., Chen, J.J., Morgan, T.A., John, M., Siegel, E.L.: Facial recognition software success rates for the identification of 3D surface reconstructed facial images: implications for patient privacy and security. J. Digit. Imaging **25**(3), 347–351 (2012)
37. Schwarz, C.G., Kremers, W.K., Therneau, T.M., Sharp, R.R., Gunter, J.L., Vemuri, P., Arani, A., Spychalla, A.J., Kantarci, K., Knopman, D.S., Petersen, R.C.: Identification of anonymous MRI research participants with face-recognition software. N. Engl. J. Med. **381**(17), 1684–1686 (2019)
38. Karimian, G., Petelos, E., Evers, S.M.: The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. AI Ethics. **28**, 1–3 (2022)
39. SERIES BP. Opportunities and Challenges of Blockchain Technologies in Health Care. Available online: https://www.oecd.org/finance/Opportunities-and-Challenges-of-Blockchain-Technologies-in-Health-Care.pdf. Accessed 26 Apr 2022
40. Mammen PM. Federated learning: opportunities and challenges. arXiv preprint arXiv:2101.05428. 2021 Jan 14.
41. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. Vol. 27. (2014)
42. Fernandez-Quilez A, Larsen SV, Goodwin M, Gulsrud TO, Kjosavik SR, Oppedal K. Improving prostate whole gland segmentation in t2-weighted MRI with synthetically generated data. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1915–1919. IEEE. (2021).
43. Fernandez-Quilez, A., Parvez, O., Eftestøl, T., Kjosavik, S.R., Oppedal, K.: Improving prostate cancer triage with GAN-based synthetically generated prostate ADC MRI. In: Medical Imaging 2022: Computer-Aided Diagnosis, vol. 12033, pp. 422–427. SPIE (2022)
44. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In International conference on machine learning. pp. 1597–1607. PMLR. (2020)

45. Borji A. Pros and Cons of GAN Evaluation Measures: New Developments. arXiv preprint arXiv:2103.09396. (2021 Mar 17).
46. Johnson, S.: Racing into the fourth industrial revolution: exploring the ethical dimensions of medical AI and rights-based regulatory framework. AI Ethics. **23**, 1–6 (2022)
47. Article 29 Data protection working party. Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679.
48. Bleher, H., Braun, M.: Diffused responsibility: attributions of responsibility in the use of AI-driven clinical decision support systems. AI Ethics. **24**, 1–5 (2022)
49. Richardson, J.P., Smith, C., Curtis, S., Watson, S., Zhu, X., Barry, B., Sharp, R.R.: Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit. Med. **4**(1), 1–6 (2021)
50. Musbahi, O., Syed, L., Le Feuvre, P., Cobb, J., Jones, G.: Public patient views of artificial intelligence in healthcare: a nominal group technique study. Digit. Health. **7**, 20552076211063680 (2021)
51. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225. (2017 Nov 14).
52. Kaplan, S., Handelman, D., Handelman, A.: Sensitivity of neural networks to corruption of image classification. AI Ethics. **23**, 1 (2021)

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