

Deep Learning and Medical Image Analysis: Epistemology and Ethical Issues

Francesca Lizzi¹^a, Alessandra Retico¹^b and Maria Evelina Fantacci²^c

¹*National Institute for Nuclear Physics, Pisa Division, Pisa, Italy*

²*Department of Physics, University of Pisa, Pisa, Italy*

Keywords: Deep Learning, Ethics, Epistemology, Multi Disciplinary Science, Medical Imaging.


Abstract: Machine and deep learning methods applied to medicine seem to be a promising way to improve the performance in solving many issues from the diagnosis of a disease to the prediction of personalized therapies by analyzing many and diverse types of data. However, developing an algorithm with the aim of applying it in clinical practice is a complex task which should take into account the context in which the software is developed and should be used. In the first report of the World Health Organization (WHO) about the ethics and governance of Artificial Intelligence (AI) for health published in 2021, it has been stated that AI may improve healthcare and medicine all over the world only if ethics and human rights are a main part of its development. Involving ethics in technology development means to take into account several issues that should be discussed also inside the scientific community: the epistemological changes, population stratification issues, the opacity of deep learning algorithms, data complexity and accessibility, health processes and so on. In this work, some of the mentioned issues will be discussed in order to open a discussion on whether and how it is possible to address them.


1 INTRODUCTION


Machine and deep learning methods applied to medical images are proving to be a promising way to improve the performance in solving many issues: the diagnosis of a specific disease, the contouring of organs or lesions and the prediction of the prognosis. Deep learning, in particular, offers the possibility of analyzing many patients' data in a reproducible way and they can be applied to carry out follow up and radiomic studies. In particular, the advent of Deep Learning (DL) algorithms in the field of medical image analyses is leading to a change in supporting physicians in their role. Many different applications have been explored (Litjens et al., 2017) successfully. However, developing a DL algorithm with the aim of applying it in clinical practice is a complex task which should take into account the context wherein the software is developed and should be used. In 2021, the World Health Organization (WHO) published the first report about the ethics and governance of Artificial

Intelligence (AI) for health (WHO, 2021). In that report, it has been stated that AI may improve healthcare and medicine all over the world only if ethics and human rights are a main part of its development. WHO recognizes that ethical guidance based on the shared perspectives of the different entities that develop, use or oversee such technologies is critical to build trust in these technologies, to guard against negative or erosive effects and to avoid the proliferation of contradictory guidelines. Involving ethics in technology development means to take into account several issues.

In this work some of the most interesting ones will be discussed. First, understanding how scientific method is changing should be at least taken into account and discussed, when developing a medical software. In fact, most of the ethical issues related to the application of DL algorithms to clinical practice are directly connected to the shift of the scientific paradigm brought by the intense use of data and data mining. Second, the way we collect data and build data sets is crucial to develop fair AI-based algorithms. This means to correctly perform the population sampling in order to prevent social biases and to preserve technical information to avoid technolog-

^a <https://orcid.org/0000-0003-0900-0421>

^b <https://orcid.org/0000-0001-5135-4472>

^c <https://orcid.org/0000-0003-2130-4372>

ical biases. In fact, the development of AI-based algorithms should pay attention to the process of image production which includes manufacturers, acquisition parameters and also the interactions with physicians. Moreover, since we should always compare the predicted results with a ground truth, a labeled data set collection should consider the inter- and intra-observer variability as well as the risk of containing a not negligible label noise (see fig. 1). In this con-

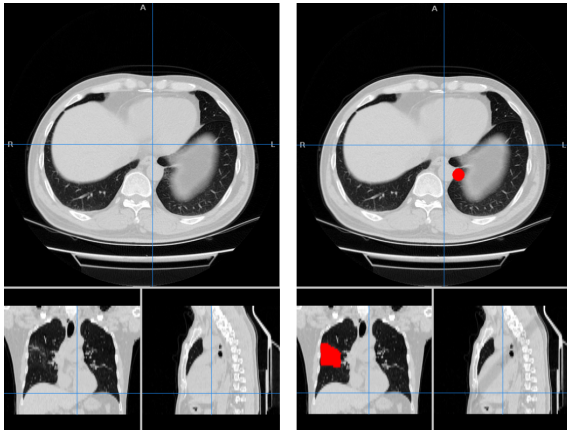


Figure 1: An example of label noise: on the left, an original lung CT scan (case 0053) from the COVID-19 Challenge data set (An et al., 2020) is shown with a windowing between -1000 and 300 HU. On the right: the reference label mask for a COVID-19 lung lesion published within this data set is shown in red. The label appears in the axial view (top row) as a perfect circle probably due to the use of a support software for image labelling by the radiologist.

text, the application of AI to medical images needs a special care since its wrong use may harm not only people but also the healthcare systems (Stevens et al., 2018).

In this work, the first section is dedicated to the changing scientific paradigm and on how the so-called Fourth Paradigm affects studies that concern deep learning and medical images. In the second section, the many issues connected to the collection of a data set are discussed, including the problem of the ground truth reliability, of the access to data and their quality, and the problem of improving the clinical outcomes in the context of a real hospital protocol.

2 THE IMPORTANCE OF EPISTEMOLOGICAL CLAIMS IN HEALTHCARE

The concept of "scientific paradigm" was introduced by Kuhn in the "Structure of Scientific Revolutions" published in 1962 (Kuhn, 1962) and, despite its lim-

its, it is very useful to frame a simplified scheme of the evolution of the scientific paradigms reported by (Hey et al., 2021), as shown in (Table 1). According to Kitchin (Kitchin, 2014), we can summarize the idea of the paradigm according to Kuhn as an accepted way of interrogating the world and produce knowledge which is common to a substantial proportion of researcher in a discipline at any one moment in time.

According to Hey (Hey et al., 2021), epistemology is moving towards a new paradigm called the "fourth paradigm" or "exploration science". In this evolution, some fundamental rules of traditional science are deeply changing and they should be taken into account since they are useful to establish the limits and the possibilities of these new rising methods. The assumptions that are made during the development of an algorithm are critical to define the model itself and the boundaries in which it should be applied. According to Gray (Hey et al., 2021), there are two main ways to frame the fourth paradigm: the first way, typical of industry, is a pure empiricism wherein machine learning techniques can reveal the inherent truth of data, while the second one looks at the fourth paradigm as a new extension of the established scientific paradigm. The fourth paradigm, as pure inductive empiricism brought by the use of Big Data and also by the rising of the deep learning methodologies, has the potential to undermine the scientific legitimacy of the machine learning (Enni and Herrie, 2021). As an example, Campolo and Crawford (Campolo and Crawford, 2020) use the Enchantment theory of Max Weber to describe a broader epistemological diagnosis of modernity. They affirm that not understanding the motivation that leads a deep learning based model to a decision could produce the effect of considering that algorithm as something magical. These considerations do not come only from humanities but also from "hard" sciences. Stuart J. Russell, a well-known professor of computer science from Berkeley University, in 2018, spoke about deep learning and described it as "a kind of magic" since we cannot understand when and why the deep learning hypothesis is consistent. For this reason, it is interesting to discuss the process of knowledge generation, evidence and causation in particular in the healthcare domain. In Stevens et al. (Stevens et al., 2018), a critically and healthcare centered review of epistemological claims is presented. The healthcare field is characterized by an institutionalized set of epistemological principles and generally accepted scientific methodologies (Stevens et al., 2018) (Beltran, 2005) which are challenged by the deep learning practices. The language used to describe the applications of algorithm in healthcare can be an interesting way of analyzing whether there are

Table 1: The evolution of scientific paradigm according to Hey et al. (Hey et al., 2021).

Paradigm	Method	Dating
Experimental Science	Observation of natural phenomena	Pre-Renaissance
Theoretical Science	Using models, generalization	Pre-Computers
Computational Science	Simulations	Pre-Big Data
Exploratory Science	Data mining	Now

different ways of using data in this specific domain. Stevens et al., studying systematically the editorials on the use of Big Data practices in healthcare, describes five ideal typical discourses, naming them using the relations between implicit assumptions about evidence and knowledge and the diverse epistemological positions. The five categories they design are: the modernist, the instrumentalist, the pragmatist, the scientist and the critical-interpretative. Despite the details of each discourse, there is a significant difference between the firsts four and the last one about the conceptualization of data: the former consider data inherently truthful and meaningful (natural and pre-existing), while the latter consider data as constructed and hence they necessarily emphasize some aspects and leave out others. While a simple positivistic, hypothesis-and-theory free and purely empiricist approach seems to be a way of making the use of data simpler in the clinical practice, it is a trap. With regard to the field of medical image analysis and consequently the radiological medical domain, we know that data are not given, natural or pre-existing. Medical images are the results of:

1. A traditional scientific process: their production is based on physical studies on the interaction between matter and radiation, human body and radiation, on the physical processes of X-ray, magnetic fields and ultrasounds production as regards radiology, and radioactivity and all the issues linked to it as regard nuclear imaging;
2. A technological development history: the image production deals with the detectors improvements, the materials used for detecting photons, the electronics which, simplifying, determine for example the spatial resolution, the contrast and other image quality characteristics. When we deal with 3D images such as Computed Tomography (CT) scans we should consider the image reconstruction algorithms which are a mix of traditional scientific research, especially mathematical one (for example Radon transform or Fourier signal analysis) and pure technological improvements such as the sliding contacts.
3. An industrial process: medical imaging systems are not equally distributed around the world and their production is highly costly producing as a

result that there are few vendors that deals with the imaging machinery market. Moreover, as a result of an industrial process, some parts of the medical images production are protected by patents which inhibits the complete knowledge on how an image is produced;

4. A function-based process: medical images are made on the basis of their utility and improved following their possible uses in hospitals. The choice of using a specific imaging modality depends on the scope (morphology or functionality and diagnosis, follow up, radiotherapy planning, ...) and on the part of the body that needs to be imaged. They are made to be presented to physicians in a way that medical doctors can interpret and taking into account the specific medical formation process they attended. Moreover, contrarily to natural images, most of medical imaging modality implies the delivering of a radiation dose to the patients, making their use a dynamic equilibrium between costs and risks, in terms of capital and health, and benefits.

For all these reasons, it is unacceptable to consider medical images data as pre-existing or natural. Moreover, applying Big Data techniques, such as machine and deep learning, cannot be considered as a free hypothesis science. Even if the hypothesis is a complex hypothesis and it is very far from the pre-Renaissance way of formulating it, we are always assuming that, given the constructed data and the context, there is a model which may solve the given task we want to study. This means that we are using data that contain already the solution. Characterizing the machine and deep learning techniques as comprehensive and intrinsically unbiased can be misleading rather than helpful in shaping scientific as well as public perceptions of the features, opportunities and dangers associated with data-intensive research (Leonelli, 2014). Finally medical images can rarely be considered as Big Data and hence the application of techniques developed on them should be even more careful as regards, for example, the generalization goal. What is at stake is our ability to produce knowledge not only in a traditional scientific way but making it from a critical position, avoiding the accusation of practising "magic" or "alchemy". It deals with knowing and assuming how much complex is to create algorithms,

especially deep learning one, with the scope of applying them in hospitals and, by assuming it, proceed towards a fair, scientific, active and impacting application of machine and deep learning to medicine.

3 BUILDING MEDICAL IMAGE DATA SETS

As discussed in the previous section, medical data sets of images cannot be considered inherently truthful as natural resources. It is interesting to discuss the issues that should be considered in a data set collection that demonstrates the thesis of the first section. In particular, the ground truth production and accessibility and quality of medical images data sets will be considered in the following.

3.1 Ground Truth Reliability

The ground truth on medical images is usually made by medical doctors opinions or by a consensus among them. For this reason, it always suffers from a certain grade of variability which should be kept under control (Bridge et al., 2016) (Renard et al., 2020). The use of a peculiar imaging modality and of a specific imaging system may affect the capability of having a reliable ground truth and the aggregation of data coming from different sources is a challenge that still need to be addressed. In fact, publicly available data sets of medical images usually are small data sets that need to be aggregated in order to obtain a set with a sufficient number of samples to train a deep learning algorithm (Lizzi et al., 2021). Even if it is possible to collect private image data sets from hospitals, the process is very time consuming for both the collection and the labeling. Moreover, the publication of such data sets may not be possible, thus reducing the chance of reproducing results obtained by other studies. Publishing the data is not easy because database maintenance is expensive and the privacy of patients has to be managed rightfully. The ground truth usually depends on the task we want to solve and its creation is a pivotal step for algorithm development. If the task we want to solve is a classification task, the ground truth consists in assigning a class to each image or patient in the data set. A patient image taken at different time points may belong to a different class because human body changes over time. The way the classes are defined is mainly based on medical protocols but medical protocols do not guarantee the variability delete. Even if this labelling process seems to be very fast, when a huge amount of labelled data is required, the process is very time consuming for doctors. Another way for

labelling medical images is to assign to each pixel or each voxel a certain class. This kind of labelling is suited for solving segmentation problems. A medical image usually contains many pixel and voxel and this characteristic makes the labelling extremely time consuming. If we suppose to have a standard lung CT scan with size 512x512x100 the number of elements to be labelled is more than 26 millions! There are some tools to help physicians in this task but they may introduce a bias in the labelling. In order to reduce the cost of labelling, the use of non-expert people has been employed in the field of natural images (Kuznetsova et al., 2020), but the use of such kind of labelling process leads to highly noise data sets. In the medical images domain, in which the objects to be identified are usually small and are difficult to be identified, this process is even harder. Having the availability of large labeled data sets of medical images is currently a real challenge even despite the labelling process. Medical images data sets, in fact, are usually small and their collection is not easy because of privacy issues and institutional policies.

3.2 Data Accessibility and Quality

Data may come from public or private collections. Both of these two modalities have weaknesses and strengths which are going to be discussed in the following. First of all, public data may be effectively public, i.e. accessible to every one, or they may be accessed through a specific agreement. Private data are instead data which can not be used or accessed in any case. Data may be not accessible for many reasons. One of the most problematic is related to privacy. In order to better understand the risks of publishing data, it is interesting to discuss the most used image formats. This is important because medical image formats usually contain a header with patient and physician information. There are mainly two data formats typically used for medical images and they are the Neuroimaging Informatics Technology Initiative (NIfTI) and the Digital Imaging and Communications in Medicine (DICOM) (Standard DICOM, 2021). The NIfTI format was created in the field of neuroimaging and it is a standard which contains a header with only information about orientation, voxel size and image visualization. 3D images, for example CT scans or MRI scans, can be stored in this format which defines uniquely the correct orientation and the physical volumes. In Figure 2, an example of a NIfTI header of a CT scan is reported.

The Digital Imaging and Communications in Medicine (DICOM) standard (Standard DICOM, 2021) is the global convention used by manufacturers


```

File Format: NIFTI
-----
Dim Info = 0
Image Dimensions (1-8): 3, 512, 512, 301, 1, 1, 1, 1
Intent Parameters (1-3): 0.0, 0.0, 0.0
Intent Code = 0
Datatype = 4
Bits Per Voxel = 16
Slice Start = 0
Voxel Dimensions (1-8): -1.0, 0.81055, 0.81055, 1.0, 1.0, 1.0, 1.0, 1.0
Image Offset = 352
Data Scale (Slope, Intercept): Slope = 1.0 Intercept = 0.0
Slice End = 0
Slice Code = 0
Units Code (Spatial, Temporal) = 0
Display Range (Max, Min): Max = 0.0 Min = 0.0
Slice Duration = 0.0
Time Axis Shift = 0.0
Description: ""
Auxiliary File: ""
Q-Form Code = 0
S-Form Code = 2
Quaternion Parameters: b = 0.0 c = 1.0 d = 0.0
Quaternion Offsets: x = 213.10715 y = -206.89215 z = -220.05334
S-Form Parameters X: -0.81055, 0.0, 0.0, 213.10715
S-Form Parameters Y: -0.0, 0.81055, 0.0, -206.89215
S-Form Parameters Z: 0.0, -0.0, 1.0, -220.05334
Intent Name: ""
-----
    
```

Figure 2: An example of a whole NIFTI header. It can be noticed that this header includes only information about the voxel size and the image orientation.

to define and store diagnostic imaging data. DICOM images are encoded as a set of elements; public elements are defined by the DICOM standard, and private elements are defined on an individual basis by each manufacturer. A DICOM data element or attribute is made of 3:

- a tag that identifies the attribute, usually in the format (XXXX, YYYY) with hexadecimal numbers, and may be divided further into DICOM Group Number and DICOM Element Number;
- a DICOM Value Representation (VR) that describes the data type and format of the attribute value.

The fields of the DICOM header contain many information from the patient ID, which is a number that uniquely identifies the patient, the Patient's Birth Name (0010,1005), the Patient's Age (0010,1010), the Patient's Size (0010,1020), the Patient's Address (0010,1040) or even the Patient's Mother's Birth Name (0010,1060). All these data are a problem when we deal with privacy because they may allow a complete re-identification of subjects. On the other hand the DICOM format contains also information on the acquisition parameters such as the reconstruction kernel, the imaging system used, exposure time, X-ray tube current, the field of view (FOV) size or the reconstructed FOV.

These characteristics are less prone to be problematic with regard to privacy and they are very useful for algorithms development. However, in most of published medical images data all this information is lost. This is mainly due to the fact that it is not so easy to

```

File Format: DICOM (CT Image Storage)
-----
Filename: 000038.dcm
(0002,0000) Group 0002 Length [206]
(0002,0001) File Meta Information Version [(2 Bytes of raw data)]
(0002,0002) Media Stored SOP Class UID [1.2.840.10008.5.1.4.1.1.2]
(0002,0003) Media Stored SOP Instance UID [1.3.6.1.4.1.14519.5.2.1.6279.6001.315606855383999143703852453142]
(0002,0010) Transfer Syntax UID [1.2.840.10008.1.2.1]
(0002,0012) Implementation Class UID [1.3.6.1.4.1.22213.1.143]
(0002,0013) Implementation Version Name [0.5]
(0002,0016) Source Application Entity Title [POSDA]
(0008,0005) Specific Character Set [ISO_IR 100]
(0008,0008) Image Type [ORIGINAL,PRIMARY,AXIAL]
(0008,0016) SOP Class UID [1.2.840.10008.5.1.4.1.1.2]
(0008,0018) SOP Instance UID [1.3.6.1.4.1.14519.5.2.1.6279.6001.315606855383999143703852453142]
(0008,0020) Study Date [20000101]
(0008,0021) Series Date [20000101]
(0008,0022) Acquisition Date [20000101]
(0008,0023) Content Date [20000101]
(0008,0024) Overlay Date [20000101]
(0008,0025) Curve Date [20000101]
(0008,002A) Acquisition DateTime [20000101]
(0008,0030) Study Time []
(0008,0032) Acquisition Time []
(0008,0033) Content Time []
(0008,0050) Accession Number []
(0008,0060) Modality [CT]
(0008,0070) Manufacturer [GE MEDICAL SYSTEMS]
(0008,0090) Referring Physician's Name []
(0008,1090) Manufacturer's Model Name [LightSpeed Plus]
    
```

Figure 3: An example of a part of a DICOM header. Contrarily to the NIFTI one, the DICOM header contains information about the patient, the instrumentation, the acquisition protocol and the clinical personnel.

treat privacy and DICOM standard, since the number of tags that may be contained is very large. Moreover, making studies on humans implies not only privacy related issues but ethical issues too. For these reasons accessing to Italian hospital data requires a strict protocol to be carried out. The image instrumentation manufacturers use private elements to encode acquisition parameters that are not yet defined by the DICOM standard or that they consider proprietary. They also define and include private elements that contain Protected Health Information (PHI). These PHI private elements can be as obvious as the name of a patient and as subtle as an identifier string that could be tracked back to a patient by someone with access to the departmental image archive. A DICOM conformance statement is a document published by a manufacturer that contains technical information concerning data exchange with a specific type of device (e.g. an imaging unit, workstation, printer, image archive). The conformance statement provides the mechanism for a manufacturer to publish the set of private elements that are stored in the DICOM files created by an imaging system. Manufacturers do not document and publish all of their private elements. For these reasons, the de-identification process should meet two conflicting requirements: (i) any PHI must not be included in exported data and (ii) the system must retain all data that describe the acquisition, such as physical parameters for individual images, as well as other parameters such as series description. De-identifying a

DICOM image is a challenging task that carries the risk of leaving in the header PHI or meta-data that makes the re-identification possible. On the other hand, the NIFTI format has been invented to have not patient information in the header but it does not allow to store important technical parameters. It could be interesting to study a new image format standard suitable for AI and deep learning algorithms which contains all the technical information while keeping the privacy risk as lower as possible. The availability of acquisition parameters is crucial to have a high quality data set and the research of a good trade-off between accessibility and quality is an urgent challenge to be tackled.

3.3 How to Gain Impact?

Over the last 10 years, publications on AI in radiology have increased from 100–150 per year to 700–800 per year (Pesapane et al., 2018) and the interest in the medicine field is continuously increasing. AI and deep learning studies mainly focus their scopes on increasing the accuracy of diagnosis when compared to the physicians performances.

However high accuracy does not necessarily mean that an AI algorithm improves clinical outcomes. It is, in fact, important to assess whether its use in clinical practice can be integrated in the hospital workflow and how much the impact is, not only on the outcomes, but also on the physicians' training. In order to perform this kind of analysis, a clinical trial is needed and clinical trial studies are usually time consuming and expensive. It is pivotal to question which could be the role of AI in the medical and clinical workflow, especially in the radiology field which seems to be the most explored field of medicine. It is also interesting to discuss the role of a radiologist in the hospital workflow and whether they can be replaced by an artificial intelligence or be supported by it. In Pesapane et al. (Pesapane et al., 2018), a group of radiologists reflects on what it means to let an AI make a diagnosis and what are the differences between the human evaluation and the AI one.

AI and especially deep learning functioning in radiology is based on a principle that is very similar to the clinical one: “the more images you see, the more examinations you report, the better you get” and this may be the reason why AI is successfully applied to radiology. Since the comparison between the radiologist's and AI performance depends on the radiologist experience and also on the quality of the developed AI, it is not straightforward to state whether and when one is better than the other. When image analysis takes too much time with respect to the neces-

sity of the patient, i.e. a very urgent clinical evaluation is necessary, AI may be very helpful in a hospital workflow. As an example, in this study (Kim et al., 2021), the application of a deep learning-based assistive technology in the Emergency Department (ED) context has been studied on Chest Radiographs (CRs). CR interpretation is a difficult task that requires both experience and expertise because various anatomical structures tend to overlap when captured on a single two-dimensional image, different diseases may have a similar presentation and specific diseases may be present with different characteristics. For these reasons, the CR interpretation suffers from a significant possibility of misinterpretation, reaching the 22% according to (Donald and Barnard, 2012). ED physicians perform worse than trained radiologists in reading images. However, radiologists may not be available, especially during nights and weekend and CR interpretation in the ED settings is given to ED physicians. For all these reasons, Kim et al. (Kim et al., 2021) studied whether an ED physician supported by a deep-learning based algorithm for CR interpretation performs better than the single ED physician. They found that ED departments may benefit from the use of AI even if this experiment needs at least an external validation study. This is an example that shows clearly how much it is important to know the healthcare domain and practice in order to structure a deep learning experiment.

Despite the improvements deep learning may produce to healthcare, another pivotal question concerns the problem of accountability. When an AI is used to make a decision in clinical practice, it is not trivial to understand who is responsible for the diagnosis. In this work (Neri et al., 2020), a radiologist supported by an AI is depicted as responsible for the diagnosis if they are trained on the use of AI since they are responsible for the actions of machines. Moreover, it is necessary to deepen the research field of explainability in order to let the radiologists understand the AI behaviour. Furthermore, the use of AI may bias the radiologist decisions. Lastly, even the public discussion on the introduction of AI systems as possible substitutes of the physicians themselves can be dangerous and produce a paradox effect: since radiologists are going to be replaced by AI, there will be a lack of motivation for young doctors to pursue a career in radiology.

For all these reasons, building and even bringing in the public debate deep learning models to be applied to radiology is a delicate task and claiming clinical advancements without clinical trials is unfair. As described above, medicine and healthcare fields are complex domains and the advancement that comes

from the introduction of deep-learning based technologies should be carefully evaluate inside that context to evaluate the real impact. Including information on context at the beginning of the development of a DL algorithm could help to gain clinical impact.

4 CONCLUSIONS AND DISCUSSION

Some of the many aspects that deal with the creation of a deep learning algorithm applied to medicine and, specifically, to imaging have been discussed. The difficulty of taking into account all the issues is clear and they relate to many fields of knowledge.

In section 2, the changing scientific paradigm has been discussed. How researchers pose their research questions and which epistemological assumptions they embrace are fundamental to understand the kind of research they are doing. This process cannot be done without looking at the social processes that leads to the data collection and the data generation.

In section 3.2, the process of collecting a data set and subsequently building a ground truth on medical images is discussed with potentialities and limitations. Typically, the ground truth on medical images is made by the physician opinion or by a consensus among many medical doctors. When made with the second modality, the ground truth always suffers from a variability that is difficult to be erased. The quality of an algorithm strictly depends on the quality of the ground truth but having a large number of physicians is economically expensive and requires a high grade of coordination and collaboration among research and health institutes. The quality of the algorithms depends also on the quality of data that can be private or public. Public data guarantees the possibility of testing different algorithms on the same data set but, in order to make them publishable, important information on, for example, acquisition protocols or scanners, may be lost. Private data has the advantage to be designed for the specific experiment and taken following inclusion criteria decided by the collector. When released, this kind of data can be designed to contain the information on acquisition that could be useful and meaningful for the analysis. In any case, medical images data are scarce and they may lack of label quality. This issue is one of the most limiting in deep learning algorithm development. Lastly, when an algorithm is developed to be used in clinical practice, it has to be validated not only statistically but also clinically. The performances on a test set are not sufficient to claim for clinical advancements since it belongs to the same data set used for the training

and the validation. The algorithm, in fact, needs to be tested also on at least an independent external data set to evaluate its generalization capability. The external data should be taken from another medical center and should contain the information on acquisition and scanners in order to make possible the analysis of the image characteristics that may confuse the algorithm. This process can be done in two modalities, case-control and clinical trial studies, and both of them suffer from the issues to correctly represent the population.

Once the algorithm has been externally validated, it should be integrated in the hospital workflow and its performance should be evaluated also in this context. It has been established that the capacity of an algorithm to outperform a physician is strictly connected to the experience of the physician to solve that specific task. For this reason, there exist situations in which the application of an algorithm may be really helpful to both increase performance and save time. In this context, it is interesting to question who is responsible for the diagnosis when an algorithm is used to support physicians or directly to diagnose a certain disease. In order to solve this issue, we need juridical instruments that helps the application of algorithms in clinical practice. Building responsibility means also to train physicians to the use of AI in order to make them mindful of its use and to produce an informed consent that patients can really understand.

All these issues suggest that the development of an algorithm for clinical applications need a deep and widespread knowledge in all of the cited fields: medicine, radiology, healthcare processes, laws, computer science, computer engineering, physics, social sciences, philosophy and so on. We believe that the future research on the mentioned issues should proceed along two paths: on one hand, we need to think and implement controlled experiments in order to systematically understand, for example, how the physical parameters of images, such as kVp or reconstruction algorithms, affects the feature learning of DL models; on the other hand, we should study how to build high quality shareable inclusive privacy-preserving data sets, not only as a benchmark for performance comparison but also for the whole algorithm development. This could be done, for example, by creating a new medical image standard that considers the use of data for scientific analysis and purpose as an intrinsic property of the standard itself.

What is at stake is to develop a high-performance, inclusive and trustworthy AI.

REFERENCES

- An, P., Xu, S., Harmon, S. A., Turkbey, E. B., Sanford, T. H., Amalou, A., Kassin, M., Varble, N., Blain, M., Anderson, V., Patella, F., Carrafiello, G., Turkbey, B. T., and Wood, B. J. (2020). CT Images in COVID-19.
- Beltran, R. A. (2005). The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care. *Journal of the National Medical Association*, 97(1):110.
- Bridge, P., Fielding, A., Rowntree, P., and Pullar, A. (2016). Intraobserver Variability: Should We Worry? *Journal of Medical Imaging and Radiation Sciences*, 47(3):217–220.
- Campolo, A. and Crawford, K. (2020). Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society*, 6:1–19.
- Donald, J. J. and Barnard, S. A. (2012). Common patterns in 558 diagnostic radiology errors. *Journal of Medical Imaging and Radiation Oncology*, 56(2):173–178.
- Enni, S. A. and Herrie, M. B. (2021). Turning biases into hypotheses through method: A logic of scientific discovery for machine learning. *Big Data and Society*, 8(1).
- Hey, T., Tansley, S., and Tolle, K. M. (2021). *Fourth Paradigm*.
- Kim, J. H., Han, S. G., Cho, A., Shin, H. J., and Baek, S.-E. (2021). Effect of deep learning-based assistive technology use on chest radiograph interpretation by emergency department physicians: a prospective interventional simulation-based study. *BMC Medical Informatics and Decision Making*, 21(1):1–9.
- Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data and Society*, 1(1):1–12.
- Kuhn, T. S. (1962). *The structure of Scientific Revolution*, volume I,II.
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. (2020). The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 128(7):1956–1981.
- Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data and Society*, 1(1):1–11.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(December 2012):60–88.
- Lizzi, F., Brero, F., Cabini, R. F., Fantacci, M. E., Piffer, S., Postuma, I., Rinaldi, L., and Retico, A. (2021). Making data big for a deep-learning analysis: Aggregation of public COVID-19 datasets of lung computed tomography scans. *Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021*, (Data):316–321.
- Neri, E., Coppola, F., Miele, V., Bibbolino, C., and Grassi, R. (2020). Artificial intelligence: Who is responsible for the diagnosis? *Radiologia Medica*, 125(6):517–521.
- Pesapane, F., Codari, M., and Sardanelli, F. (2018). Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, 2(1).
- Renard, F., Guedria, S., Palma, N. D., and Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. *Scientific Reports*, 10(1):1–16.
- Standard DICOM (2021). DICOM standard.
- Stevens, M., Wehrens, R., and de Bont, A. (2018). Conceptualizations of Big Data and their epistemological claims in healthcare: A discourse analysis. *Big Data and Society*, 5(2):1–21.
- WHO (2021). *Ethics and Governance of Artificial Intelligence for Health Ethics and Governance of Artificial Intelligence for Health 2*.