Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Machine Learning for Clinical Decision-Making: Challenges and Opportunities

Sergio Sánchez-Martínez [1], Oscar Camara [2], Gemma Piella [2], Maja Cikes [3], Miguel Ángel González Ballester [2-7], Marius Miron [4], Alfredo Vellido [5], Emilia Gómez [2-4], Alan Fraser [6], Bart Bijnens [1-7]

1 August Pi i Sunyer Biomedical Research Institute (IDIBAPS), Barcelona, Spain
2 University Pompeu Fabra, Department of Information and Communication Technologies, Barcelona, Spain
3 University of Zagreb School of Medicine, Department of Cardiovascular Diseases, Zagreb, Croatia
4 Joint Research Centre, European Commission, Brussels, Belgium
5 Computer Science Department, Intelligent Data Science and Artificial Intelligence (IDEAI) Research Center, Universitat Politecnica de Catalunya, Barcelona, Spain
6 Wales Heart Research Institute, School of Medicine, Cardiff University, Cardiff, UK
7 ICREA, Barcelona, Spain

Acknowledgements: This study was supported by the Spanish Ministry of Economy and Competitiveness (“María de Maeztu” Programme for R&D [MDM-2015-0502], Madrid, Spain) and by the “Fundació La Marató de TV3” (nº20154031, Barcelona, Spain). The work of S.Sanchez-Martinez was supported by IDIBAPS and by the HUMAINT project of the Joint Research Centre of the European Commission. Special thanks to Nicolas Duchateau, who carefully read this manuscript and provided valuable feedback.

Corresponding author:

Sergio Sánchez-Martínez, PhD
Institut d'Investigacions Biomèdiques August Pi i Sunyer,
Level -1, Table #71, Carrer Mallorca, 183, 08036, Barcelona, Spain
Email: ssanchezm@clinic.cat, sersanmarsergio@gmail.com

Total word count: 10786 words.

© 2019 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Abstract

The use of machine learning (ML) approaches to target clinical problems is called to revolutionize clinical decision-making. The success of these tools is subjected to the understanding of the intrinsic processes being used during the classical pathway by which clinicians make decisions. In a parallelism with this pathway, ML can have an impact at four levels: for data acquisition, predominantly by extracting standardized, high-quality information with the smallest possible learning curve; for feature extraction, by discharging healthcare practitioners from performing tedious measurements on raw data; for interpretation, by digesting complex, heterogeneous data in order to augment the understanding of the patient status; and for decision support, by leveraging the previous step to predict clinical outcomes, response to treatment or to recommend a specific intervention. This paper discusses the state-of-the-art, as well as the current clinical status and challenges associated with each of these tasks, together with the challenges related to the learning process, the auditability/traceability, the system infrastructure and the integration within clinical processes.

Keywords: machine learning; clinical decision-making; personalized medicine; digital health

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Introduction

According to the *Independent High-Level Expert Group on Artificial Intelligence* set up by the *European Commission* (Smuha, 2019) [1], "*Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension* ***by perceiving their environment*** *through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from these data and* ***deciding the best action(s)*** *to take to* ***achieve the given goal***." For this, "*AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions*." "*As a scientific discipline, AI includes several approaches and techniques, such as* ***machine learning*** *(of which deep learning and reinforcement learning are specific examples),* ***machine reasoning*** *(which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics."*

In the context of exploiting machine learning (ML) for clinical decision-making, these definitions imply that a (software) system would **perceive** a given individual by collecting and interpreting data relevant for that individual's health, and **reason** on these data to **suggest the best actions** to take in order to maintain or improve the person's health.

This is similar to the traditional approach that a medical professional takes when examining and treating a sick patient or when suggesting preventive actions to avoid illness. Therefore, to assess the opportunities as well as the challenges of ML systems for clinical decision-making, a more detailed analysis of this process, when performed by clinicians, is appropriate. Additionally, for the clinical community to definitively embrace and feel confident with the use of ML in healthcare, the provided systems should be close to the process by which current clinical decisions are made, and follow the Hippocratic oath: "First do no harm" (Wiens et al., 2019).

Figure 1 summarises a typical paradigm for clinical decision-making. It starts by data acquisition, including gathering the clinical history of the patient, recording demographics, performing simple measurements such as weight, height, and blood pressure, acquiring an electrocardiogram, and obtaining imaging and laboratory tests, and it is accompanied by the extraction of relevant features/indices from the collected data (e.g. body mass index, image measurements, heart rate, among others). Next, clinicians construct and interpret the **state** of the patient by grouping the relevant data items and **comparing these with population-based information** learned during their (continuous) education; derived from guidelines; or with types of patients they encountered during their career. This interpretation/comparison is based on reasoning on the data using the human excellent capability of putting information into context through pattern recognition. However, although the interpretation/classification is based on the available data, it is crucial to keep in mind that clinicians implicitly take into account the **uncertainty** associated with measurements as well as the

[1] https://www.aepd.es/media/docs/ai-definition.pdf

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

completeness of the available information and consider to which amount they can **rely** on the data. Based on the comparison of the patient's state to knowledge from the (natural as well as treated) expected evolution of related populations, they aim to make **interpretable** decisions, which can explicitly be traced back to **understand the reasoning** behind them. Some of the resulting actions from made decisions can be to either collect more data to minimize the uncertainty and maximise the reliability associated with the decision, additionally considering the cost-benefit trade-off; to make an intervention (drug/device therapy, surgery, etc.) in order to improve the state of the patient and thus the outcomes; or to send the patient home (whether or not with planned observation follow-up).

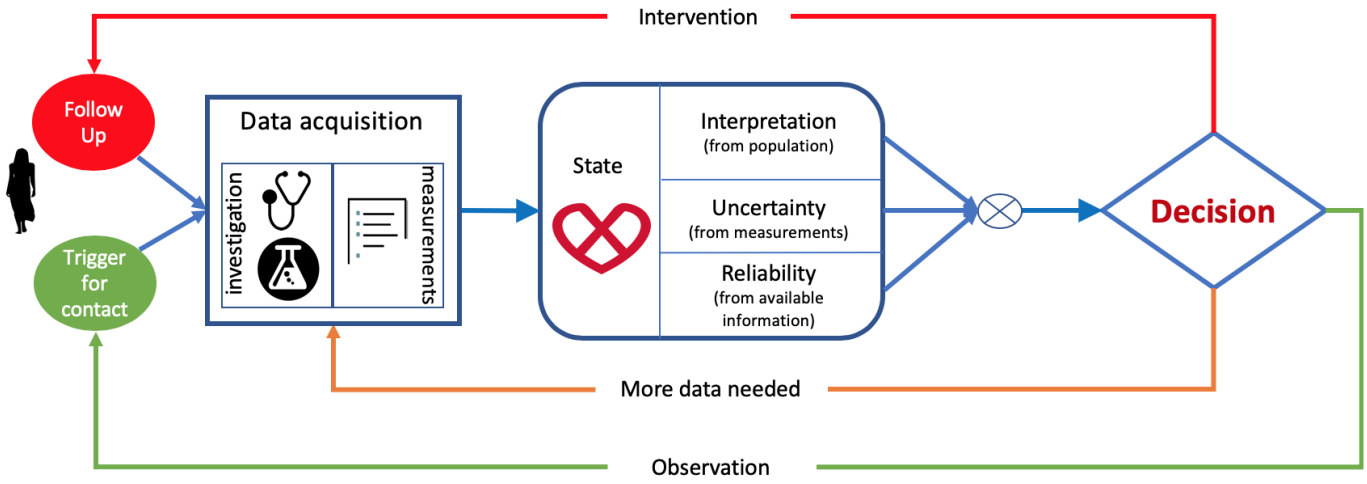

Figure 1. Clinical decision-making flowchart, from data acquisition and extraction, to patient's status interpretation and associated decision

Besides the above flowchart (often implicitly) used during clinical decision-making, it is important to keep in mind that the human brain is considered to have two distinct ways to reason on data to come to an interpretation by comparing to previous knowledge (Croskerry, 2009): the *fast/intuitive* (Type 1) versus the *slow/reasoned* (Type 2) approach. The first is almost instantaneous and based on the human ability to apply heuristics to recognise patterns; it is typically used in medical emergencies. However, using the Type 1 approach for clinical decisions is prone to error given that it can easily lead to incomplete *perception* of the patient and under-interpretation of the data (such as missing additional relevant conditions) or relying on low quality data, assuming it captures all features relevant to the patient (Benavidez, Gauvreau, & Geva, 2014; National Academies of Sciences, Engineering, 2015). On the other hand, the Type 2 approach is deductive, deliberate, and asks for a significant greater intellectual, time and cost investment, but often turns out more accurate. As each patient represents a unique composition of disease and comorbidities, to optimize decisions, we certainly need to ensure that both the reasoning process and the population of reference capture as many as possible exceptions, or at least help identifying the outliers, especially in circumstances where *fast* (Type 1) reasoning could dominate. The above is highly relevant for both "traditional" clinical decision-making and for ML-based systems, given that different algorithms mimic different human reasoning approaches and can lead to the same errors or problems.

To pursue an evidence-based personalised medicine (Consortium, 2001), current policies from professional healthcare organisations and healthcare providers as well as current practices from the industry stimulate

the collection of massive amounts of data (Wallentin, Gale, Maggioni, Bardinet, & Casadei, 2019). Consequently, millions of individuals are carefully examined, resulting in a deluge of complex, heterogeneous data. However, in addition to the fact that much of the (raw) data is not easily accessible, the challenge lies in extracting the most out of them to aid clinical decision-making in a way that is reproducible. Additionally, the decision-making system should consider that it is not always obvious which information will be relevant for a given patient at a given moment, due to the large inter-individual variations.

The use of algorithmic approaches to digest these complex, heterogeneous data and use them in the clinical-decision process has skyrocketed due to the ever-increasing computing power, and the latest advances in the ML field. Indeed, big data leveraged by ML approaches, especially when given a relevant place in the above described flowchart (Figure 1), could substantially contribute to clinical decision-making (Gulshan et al., 2016) , by providing refined/well-curated information to clinicians so they can make well-founded diagnoses and treatment recommendations, while also supplying a probability on the possible outcomes and cost for each one. ML-augmented decisions made by clinicians have the potential to improve outcomes, lower costs of care, and increase patient and family satisfaction.

This paper focuses on ML as a subfield of AI and on clinical decision-making as an essential part of medical practice. Clinical data include imaging, auditory (e.g., respiratory, speech of patients), temporal signals (heart rate, electrocardiogram, electroencephalogram), text (medical records, transcription of patients' symptoms), continuous variables (e.g., laboratory results, demographic measurements), etc. However, given that imaging is one of the areas to which ML has contributed the most (Liu et al., 2019), we emphasize the medical imaging field in our literature review. In the following, we discuss which are the essential building blocks needed to achieve the high-level task of, as well as their incorporation within, clinical decision-making, namely data acquisition, feature extraction, interpretation, and decision support (these can be identified in Figure 1). For each of these blocks, we review the ML state-of-the-art, and comment on the current penetration of ML tools into clinical practice (see *Clinical status* subsections) as well as on their challenges and opportunities. Finally, we elaborate on the general challenges that still need to be addressed.

This paper addresses potential questions arising from data scientists, industrial partners and funding institutions, helping them understand clinical decision-making and identify potential niches for their solutions to be helpful. At the same time, this review aims at informing clinicians about the ML state of the art, discussing which approaches could be used to target their questions and what are their current limitations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

ML in Clinical decision-making – different opportunities for different tasks

ML analyses have, to date, demonstrated human-like performance in repetitive, low-level tasks, where pattern recognition or perception play a fundamental role. Some examples are **data acquisition**, standardization and classification (Madani, Arnaout, Mofrad, & Arnaout, 2018; Nie, Cao, Gao, Wang, & Shen, 2016), or **feature extraction** (Desai et al., 2016; Kamnitsas et al., 2017). For higher-level tasks involving reasoning, such as patient's **status interpretation** and **decision support**, ML allows for the integration of complex, heterogenous data in the decision-making process, but these are still very immature and need substantial validation (Topol, 2019). In parallel to Figure 1, which illustrated the process of making clinical decisions, Figure 2 describes the tasks involved in this process according to how ML could contribute. The figure also illustrates that the risks to a patient from erroneous conclusions increase with each step.

Different ML approaches can complement clinicians in the different tasks involved in decision-making, especially aiding pattern recognition when complex data are considered, but they need to be interpretable and supervised by humans, who are often more aware of the context and the big picture. Some of the major challenges to be met are related to the creation of interpretable systems that also deal with measurement uncertainty (Doshi-Velez & Kim, 2017; Rudin, 2018).

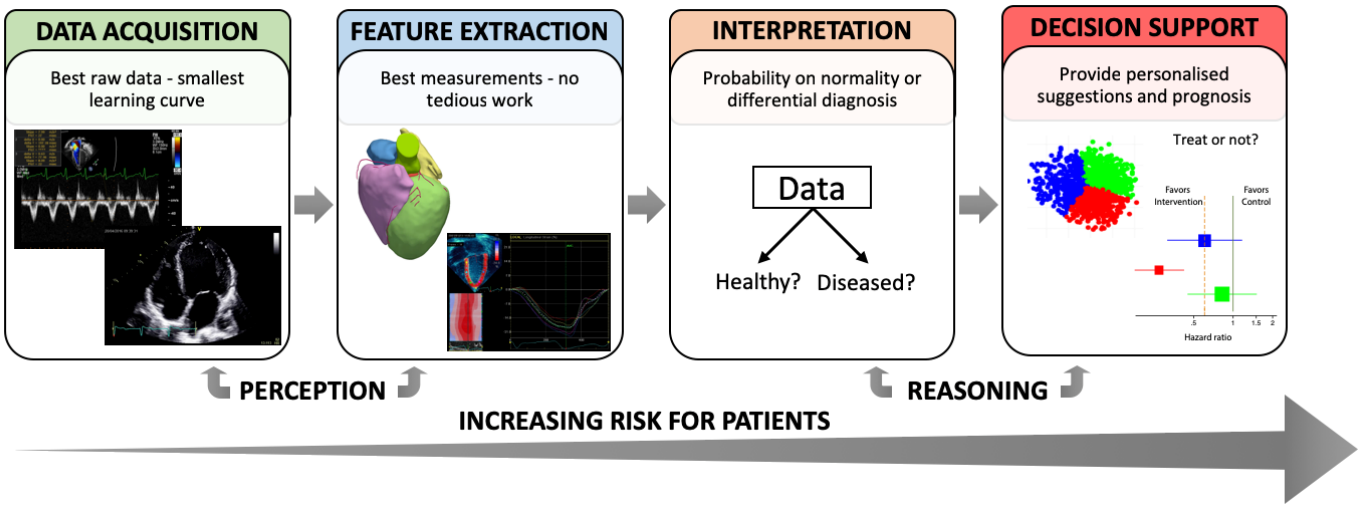

Figure 2. Different tasks where ML can support clinical decision-making.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

1. Data acquisition

Data acquisition is the first stage of the decision-making flowchart. It is of capital importance, as the extracted knowledge will be heavily conditioned by the intrinsic quality and amount of the available input data. Ideally, these data should capture the phenomenon under study, especially including relevant outliers (such as related to rare diseases or co-morbidities). To reach a decent understanding of most disease entities, a significant amount of data is required, the more representative, the better. However, this might be constrained when economic factors or invasiveness of the tests are considered.

Despite the high growth rate of healthcare data (at least 48% annually, (Minor, 2017)), they are frequently acquired using multi-vendor machinery and stored in different systems, which hinders accessibility. Moreover, medical data, such as images, are acquired by physicians/healthcare personnel with differing skill sets, time constraints and using different protocols, which precludes the straightforward combination of data from different studies and hampers the reproducibility and validity of algorithms. Here, ML (and especially deep learning (DL)) has emerged as a convenient approach to boost data quality, standardization, and reduce variability in acquisition. For data acquisition, some research deals with practicalities, e.g., DL efforts to help the operator achieving the best possible images for a given patient (Østvik, Smistad, Aase, Haugen, & Lovstakken, 2019), or classifying different imaging views (Madani, Ong, Tibrewal, & Mofrad, 2018; Østvik et al., 2019). Others focus on image reconstruction, such as AUTOMAP (Zhu, Liu, Cauley, Rosen, & Rosen, 2018), which yields higher quality data from less exposure, thus allowing to reduce radiation doses for computed tomography (CT) and positron emission tomography (PET) scans and shortening scan times for magnetic resonance imaging (MRI) (Qin et al., 2019). To do so, a deep neural network automatically figures out the transfer function of the imaging system to then come up with the correct image reconstruction model. Also, generative adversarial networks (GANs) have been proposed to generate patient images from incomplete data, or even to enrich imaging datasets for learning purposes (Kazeminia et al., 2018). However, these approaches to image reconstruction are suitable for the average patient, but might become problematic for outliers, as they rely heavily on the training dataset and often intrinsically smooth out potentially important details in the data (Kang, Min, & Ye, 2017; Zhu et al., 2018). In terms of data standardization, the bulk of current research focuses on electronic health records (Pathak et al., 2013), with some efforts geared towards data normalization (Janowczyk, Basavanhally, & Madabhushi, 2017). To aid in quality control, a collection of image generation and enhancement methods using DL have been proposed, ranging from removing obstructing elements in images (Yang et al., 2017), to improving image quality (Oktay et al., 2016; Tom & Sheet, 2017; Vedula, Senouf, Bronstein, Michailovich, & Zibulevsky, 2017; Yoon, Khan, Huh, & Ye, 2019), or completing images (Li et al., 2014). Other research focuses on controlling the quality of DL results in the context of assessing the cardiac mechanical function (Ruijsink et al., 2019).

Clinical status

Some DL applications for acquiring medical images are commercially available (WO2018220089A1, 2018), empowered by the fact that they do not need to be validated as extensively as other higher-level solutions

that involve reasoning (interpretation or decision-making). The risks associated with their use are low because they are not used to make high-level clinical decisions. To avoid propagating errors down the decision-making flowchart, data acquisition solutions already are or should be otherwise equipped with confidence estimates.

The main benefit of ML for acquiring data is that it extracts standardized, high-quality information with the smallest possible learning curve, or even without the need of human intervention (Housden et al., 2017). Advances in flexible ultrasound transducer arrays will soon allow ML to find the best combination of probes to acquire all standard echocardiographic views effortlessly (US7878977B2, 2005). Hand-held devices are already a reality in echocardiography (US20140288428A1, 2014). Together with advances in hardware, ML can be used to learn the transfer function of high-quality equipment, in order to get high-quality data with cheap devices (Yoon et al., 2019). These systems can also confirm that the images are good enough for quantification (Wu et al., 2017). The standardization of data significantly reduces inter-operator variability (Narula, Shameer, Salem Omar, Dudley, & Sengupta, 2016), which consequently reduces the variability in diagnosis. Furthermore, improved data quality will likely lead to a more precise diagnosis, as no noisy, ambiguous data is involved at the time of making decisions.

The advantages of ML for data acquisition will, in the short term, democratize access to complex acquisition systems (such as imaging techniques), bringing low-income countries within reach for quality clinical care (USAID, Rockefeller, & Gates, 2019). Coupled with efforts focused on substantially reducing the cost of scanners, the advances in data acquisition powered by ML will allow easy access to high-quality data without the need of highly trained operators. This might contribute to improving clinical diagnosis globally.

Challenges

In order to automate the extraction of data from raw imaging, a well-curated, labelled database is needed. Initial efforts are being made by the European Commission [2], who proposes to build a large-scale dataset of medical images to improve diagnosis and treatment of the most common forms of cancer. When labels are incomplete, transfer learning (Goodfellow, Bengio, & Courville, 2016)–a general purpose learning strategy that employs a model developed for a task as the starting point for a model on a second task–can be used to address this lack of data by using networks pre-trained on a generic image classification problem to further refine their connections and finally enrich the collection of labeled images for further training. Another challenge for training is inter-vendor variability, which might be overcome by showing the network enough examples of each manufacturer. When using ML-based image reconstruction with markedly reduced data sampling, the risk exists that abnormalities are masked or ‘invented’. Data compression has been used, assessed and evaluated in medicine over many years. However, in-depth validation is needed to determine the maximal data reduction allowed to still provide the current standard of information content.

[2] http://europa.eu/rapid/press-release_MEMO-18-6690_en.htm

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Table 1 organizes the ideas discussed for data acquisition in the form of a SWOT analysis, namely into Strengths, Weaknessess, Opportunities and Threats.

Table 1. SWOT analysis – data acquisition

Strengths	**Weaknesses**
• Obtain high-quality data with less exposure, cheaper devices, and minimal human intervention • Boost data quality and standardization • Enable the standardization of diagnoses • Enhance the accessibility of imaging techniques in low-income countries	• Need well-curated, representative databases for training • Need validation over time–each time the model is re-trained using new data
Opportunities	**Threats**
• Use current computing power capabilities • Use current storage capacities • Use federated learning strategies	• Miss outliers at reconstruction • Compromise diagnosis due to flawed training data • Obtain unfair results if training data is biased • Incur in domain mismatch–a ML model may fail when deployed in a context different from where it was developed

2. Feature extraction

Once the data are optimally acquired and curated (i.e. standardized and organized), the next stage towards improved clinical decision-making is to extract features for knowledge inference purposes. Machine learning techniques enable the extraction of relevant features from data such as signals (Mincholé, Camps, Lyon, & Rodríguez, 2019), lab results (Gunčar et al., 2018) or imaging, among others. In imaging, for example, these features can be either geometrical, functional or based on local voxel intensity or texture and relate to an anatomical region (organ or body substructure), whose detection is automatically achieved by segmentation, currently the most common subject of ML papers in medical imaging (Ronneberger, Fischer, & Brox, 2015). In (Moeskops et al., 2016), the authors employ a single convolutional neural network (CNN) to segment brain MRI and the coronary arteries in cardiac CT angiography. Similarly, a CNN model has been employed for segmentation of cardiac chambers across 5 common echocardiographic views (J. Zhang et al., 2018). The output was used to quantify chamber volumes, left ventricular mass and ejection fraction, and automatically determine longitudinal strain through speckle tracking. Segmentation was also used to compute tissue motion and estimate global longitudinal strain from 2D echocardiographic images, using CNNs (Østvik, Smistad, Espeland, Berg, & Lovstakken, 2018). More DL solutions in cardiovascular applications can be found in (Bernard et al., 2018; Litjens et al., 2019).

Segmentation techniques have also been geared toward lesion segmentation, often focusing on local texture. This application combines the challenges of object detection and structure segmentation. To this end, ML techniques have been used for segmentation of multiple sclerosis lesions in MRI (Brosch et al., 2016; Danelakis, Theoharis, & Verganelakis, 2018). Similar examples exist, for segmentation of traumatic brain injuries, brain tumours, and ischemic stroke (Kamnitsas et al., 2017); for segmentation of skin lesions (Vesal, Ravikumar, & Maier, 2018); for segmentation of liver lesions (Bellver et al., 2017); and for segmentation of acute ischemic lesions in diffusion-weighted imaging (Chen, Bentley, & Rueckert, 2017).

An example of a feature extracted from cardiac imaging that drives clinical decisions is left ventricular ejection fraction (Pellikka et al., 2018)–routinely used to assign treatment and predict outcomes in heart failure patients. Whereas cardiac chamber segmentation is starting to be reliably done using ML and based on different imaging modalities (Litjens et al., 2019), when quantified by humans, ejection fraction commonly shows an ordinal distribution of 5% bins, even when asked for a continuous measurement in the context of a randomised clinical trial (Kutyifa et al., 2013). Due to this, the clinical operator implicitly assumes that the measurement error is indeed in this order of magnitude and additionally incorporates knowledge on the fact that guidelines often refer to multiples of 5% when defining thresholds. This obviously poses problems for learning and integrating ML-based systems in clinical practice. Furthermore, ejection fraction often oversimplifies cardiac pathophysiology, and fails to capture the complexity of accompanying comorbidities (Cikes & Solomon, 2016). At clinical evaluation, physicians take into account these drawbacks and confer some flexibility to their decision. Indeed, clinical guidelines should presuppose that physicians flexibly adapt their diagnosis based on the reliability of measurements. The challenge for ML is to have a similar flexible behaviour by incorporating uncertainty notions (Begoli, Bhattacharya, & Kusnezov, 2019); rather than providing a firm answer, ML approaches should output confidence intervals. Likewise, the clinical interpretation of the ML results should be inherently flexible.

Uncertainty in clinical measurements causes a significant percentage of diagnostic errors (Benavidez et al., 2014). Thus, rather than generating binary diagnoses based on rigid cut-off thresholds, the ML characterization should consider continuous distributions of disease entities. To illustrate that this is a pertinent problem in clinical reality, using the ejection fraction example (Triposkiadis et al., 2019), an intermediate category of heart failure patients – Heart Failure with Mid-Range Ejection Fraction (Lam & Solomon, 2014) – has been defined. This new category illustrates that, at the time of making comparisons to support the diagnosis, we should compare individuals to entire populations rather than being limited to homogeneous cohorts confined to the strict inclusion criteria of clinical trials. Machine learning can definitely contribute to this arena by drawing conclusions from the appropriate retrospective patients.

Clinical status

The bulk of successful DL efforts in healthcare applications to date has been in feature extraction. State-of-the-art DL algorithms allow obtaining improved measurements in a time-efficient manner – e.g.,

echocardiographic studies that require half an hour to measure and reach conclusions will take seconds in the near future, thus avoiding tedious manual work. Again, inter- and intra-user variability will be reduced, paving the way for a more standardized diagnosis. In this sense, physicians will be discharged from repetitive monotonous activities such as measurement extraction, data preparation or standardization, and will be able to dedicate more time to higher-level tasks involving interpretation, patient care, and decision-making (Sengupta & Adjeroh, 2018).

In combination with radiomics (Lambin et al., 2012), i.e., the extraction of large amounts of quantitative features (e.g. shape, texture, luminosity, etc.) from images, DL can positively impact diagnosis performance (Bizzego et al., 2019; Lao et al., 2017). ML can also help discovering previously unidentified relevant features for diagnosis (Sanchez-Martinez et al., 2018). Like for data acquisition, feature extraction applications can be validated using clinical ground truth from which performance metrics are derived. Additionally, the result can often be presented in a visual way for quick inspection by the clinician, thus immediately providing feedback on the quality. Given that in many cases the extracted feature will represent only a small portion of the information used for the decision-making, the feature extraction step is associated with a low risk of leading to a wrong decision.

Challenges

As mentioned for data extraction, the lack of well-curated and sufficiently representative databases can be problematic–e.g., to successfully segment cardiac chambers in congenital heart disease cases, where the variability in morphology is far larger than within normal cases, or to detect scars in delayed-enhancement MRI. Semi-supervised learning strategies, which make use of labelled and unlabelled data for training, partially address the lack of well-curated data. In (Madani, Ong, et al., 2018), the authors employed a semi-supervised learning strategy that achieved an 80% accuracy for classifying views with only 4% labelled data and 92% accuracy for identifying left ventricular hypertrophy. Other prominent challenges when applying ML for extracting features are those associated with detecting objects and segmenting organs and substructures; if mistaken, the error will propagate and will condition the accuracy of the extracted features.

Deep learning offers the possibility of creating end-to-end models, i.e., models that represent the complete target system, bypassing the pre-processing often present in traditional approaches. These models rely less on intermediate feature extraction, thus standardizing the data pre-processing (Xue, Brahm, Pandey, Leung, & Li, 2018). However, for clinical decision-making, the opacity of this end-to-end solutions is an issue that hampers adoption.

Table 2 organizes the ideas discussed for feature extraction in the form of a SWOT analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Table 2. SWOT analysis – feature extraction

Strengths	**Weaknesses**
• Extract biomarkers that aid diagnosis • Standardize data–reduction of inter & intra-user variability • Time-efficient as compared with current clinical standards	• Need well-curated, representative databases for training • Need validation over time–each time the model is re-trained using new data
Opportunities	**Threats**
• Discharge healthcare practitioners from repetitive tasks, so they can focus on higher level activities • Discover unidentified features that may be relevant for diagnosis through radiomics • Make recommendations, but giving clinicians the opportunity to correct automatically-obtained measurements • Can be validated–use of ground truth to derive performance metrics	• Miss outliers–segmentation of rare cases • May incur in bias–gender, ethnicity, age • Depend on data quality – problematic in low-income regions

3. State interpretation – comparison to population

Once the clinical data of a patient are acquired and relevant features are extracted, the next stage in the decision process consists in interpreting his/her state by comparison to other individuals from a diseased or control population. For inter-individual comparison to be possible, data normalization is required. When complex data are involved, such as brain or cardiac images, the traditional approach to normalization is to build a statistical atlas–a reference model that captures the variability associated to the training population (Fonseca et al., 2011). To build an atlas, the training data must be transformed into a common spatio(-temporal) framework, which can be achieved by applying registration algorithms. In this sense, registration allows for the integration and efficient comparison of population data, a crucial step for state interpretation towards diagnosis.

Registration is another field where DL has shown to be beneficial. Two strategies are common: 1) using DL architectures to infer a similarity measure to drive an iterative optimization strategy to determine the optimal transformation parameters; 2) using deep regression networks to directly predict the transformation parameters. In (Dalca, Balakrishnan, Guttag, & Sabuncu, 2019), unsupervised deep feature learning is used for 3D brain registration. These authors have developed a learning framework for deformable medical image registration that promises to speed up registration time while keeping the state-of-the-art accuracy yielded by free-form deformation models (Balakrishnan, Zhao, Sabuncu, Guttag, & Dalca, 2019). In (Miao, Wang, &

Liao, 2016), Miao and colleagues used CNNs to perform 3D to 2D x-ray registration to assess the pose and location of implanted objects during surgery. Deep learning has also been used for non-rigid registration of cardiac cine-MRI sequences, achieving state-of-the-art performance while demonstrating more regular deformation fields (Krebs, Delingette, Mailhe, Ayache, & Mansi, 2019).

The interpretation of the state of the patient should be supervised by human oversight of the outcomes of machine learning, as discussed in Section “Integration – man/machine coexistence”. Indeed, a ML algorithm cleared by the FDA to automate the diagnosis of wrist fractures improved predictive values when clinicians used it, as compared to clinicians alone (Voelker, 2018). Other examples have demonstrated the synergy of man/machine by improving diagnosis of knee injuries using MRI (Bien et al., 2018).

Many other examples of patient’s status interpretation supported by ML exist, although not necessarily focused on the man-machine interaction. In (Madani, Ong, et al., 2018), DL achieved 92.3% accuracy for left ventricular hypertrophy classification (but on the other hand classified an obvious dilated cardiomyopathy as normal, thus creating the potential for a serious diagnostic mistake if not put in context). Another interesting application of DL is on predicting Alzheimer’s disease approximately 6 years before the disease strikes, using PET brain scans (Ding et al., 2018). In dermatology, a few recent examples of DL showed higher accuracy than experts at diagnosing skin cancer (Esteva et al., 2017; Haenssle et al., 2018; Tschandl et al., 2019). In cardiology, a novel DL approach was implemented to automate diagnosis of acute ischemic infarction on CT (Beecy et al., 2018). Lastly, a very recent example of DL was proposed to triage adult chest radiographs into normal or abnormal, and to grade findings as requiring non-urgent, urgent, or critical attention (Annarumma et al., 2019). Importantly, this approach achieved up to a 4-fold reduction in the average reporting delay for critical and urgent imaging findings compared with historical data.

Using information from electronic health records with a time-series analysis, clusters were identified in patients with autism spectrum disorders (Doshi-Velez, Ge, & Kohane, 2014). Another application that combined imaging data with text was used to learn semantic descriptions (labels) of radiology images from their associated reports during training, to then label a large data set containing frequent disease types (Shin et al., 2015).

Unsupervised dimensionality reduction is an example of state interpretation, a label-agnostic approach where all data are used to identify the most relevant sources of variation that describe a population, and to order all individuals according to their similarity. In these approaches, individuals with a similar clinical presentation are grouped together, whereas those showing distinct pathophysiological features are positioned far away. Although this can be used for a (continuous) quantification of a diagnostic label with different gradations of abnormality, it also provides an intuitive approach towards the assessment of therapies and interventions given that all of these are aimed to restore an individual towards increased ‘normality’. Recent examples of unsupervised dimensionality reduction provided useful insight in (large) complex patient populations (Cikes et al., 2019). Additionally, dimensionality reduction can be used to study temporally dynamic phenomena,

such as when the condition of a patient changes over a period of time as a result of an intervention (Nogueira et al., 2017).

Clinical status

The tendency of ML, and especially DL, approaches for patient's status interpretation is to mimic the idea behind content-based image retrieval, i.e., to enhance discovery in massive databases by offering the possibility to identify similar cases. This aids in building normality statistics, understanding rare disorders, and, may aid in improving patient care. Whether ML approaches are intended for diagnosis or risk assessment, they could contribute to deliver a better healthcare.

Unfortunately, many current examples of ML for interpretation of clinical data present a technically sound contribution, but often without a deep understanding of the clinical needs and many times focusing on binary classification of normal versus a specific condition, which hampers their use in routine clinical practice. Furthermore, studies showing impact on hard clinical endpoints rather than on surrogate measures are still needed. The way forward is through further integration of technical and clinical contributions, and through the elaboration of guidelines on the best way to tackle a clinical necessity using ML.

Challenges (apply to both interpretation & decision-making)

Unlike for data acquisition and feature extraction tasks, associated to a low risk of intervention with ML algorithms, applications concerning feature interpretation and decision-making imply a much higher risk, as decisions derived from them could be harmful for patients. Accordingly, ML outcomes need to be intuitively interpretable by the clinician and validated in a much more exhaustive way (as required by medical device regulators; e.g., class IIa or IIb routes to commercialization), ultimately with the launch of randomized prospective trials.

One of the main challenges for ML approaches to state interpretation lies in the extraction of robust feature representations from multidimensional input data and associating them with meaningful concepts. However, this challenge entails many others, related to the data themselves. The first is drawn from how reliable, as well as representative, the data are, both input data as well as associated outcomes. In case they were reliable, ML models need to be able to integrate these very heterogeneous data onto a single common reference space, which is not trivial. Furthermore, for the state interpretation to be successful, the data need to be drawn from a diverse-enough population, which cover gender-, ethnicity- and age-related changes, or the variability inherent to rare outliers. Data collection protocols need to be specifically designed with this objective in mind, such as in the Multi-Ethnic Study of Atherosclerosis (MESA) study (Bild et al., 2002), to minimize any of these biases. In the worst case, the system should at least be able to detect outliers or cases far from normality, even if it cannot interpret them further. On top of this, most ML systems are intended for the analysis of data collected at one timepoint, but as clinical decisions are often taken considering

longitudinal data, they should be able to assess a patient over time, e.g., during a stress protocol (Nogueira et al., 2017) or for disease (e.g. Alzheimer's Disease) progression (G. Lee, Kang, Nho, Sohn, & Kim, 2019). Finally, ML models are trained on three different kinds of data: 1) randomized clinical trials, which collect data following extensive protocols but whose strict inclusion criteria make it difficult to extrapolate from them; 2) cohorts, which are not as rich, group-balanced and standardized, and may miss some of the important parameters needed for decision-making; and 3) clinical routine real-world data, whose completeness, quality and accessibility are suboptimal, often missing clinical labels that are key for ML training. The exchange of knowledge throughout these collections of data is thus a major challenge, as what was learned from highly curated data (e.g., randomized clinical trials) will not easily generalize to routinely collected data.

Another crucial problem associated to currently available data is bias, i.e., when the training sample is not representative of the population of interest (see section "General challenges" for more details). Additionally, the subsequent learning should be fair–i.e., the population of interest should be equally treated. Often, clinical data contain some sort of bias, e.g., a clinical study performed in northern Europe will primarily enrol Caucasian individuals, and thus, ethnic minorities will be underrepresented. Consequently, what was learned in this northern European population will not be easily extrapolated to these minorities, and so the learning will be unfair. This is closely related with the statistical concept of generalizability, i.e., the ability of a model to make accurate predictions on new, unseen data that have similar characteristics as the training data. This property has been much questioned in the past and is closely related to the amount and the source of data that are used during training, and how representative these data are. Caution is needed when transferring a model learned with data from a certain clinical centre to a new one. An example of a DL model failing to detect pneumonia in chest radiographs concluded that the large differences in disease burden across sites may confound predictions (Zech et al., 2018). Other example is that of Google's ML tool to detect the cause of blindness in diabetic patients, which did not properly work during testing in rural India, arguably due to differences in data quality between training and testing data (Corinne Abrams, 2019).

Automation bias, defined as the human tendency to accept a computer-generated solution without searching for contradictory information (Cummings, 2004), may also affect clinical interpretation and decision-making. It has been shown that when the computer-generated solution is reliable, it leads to improved human performance as opposed to not having an aid. However, when the provided solution is incorrect, human error rates increased (Goddard, Roudsari, & Wyatt, 2012). Following this premise, who is to blame if a diagnostic algorithm errs at spotting a cancerous nodule on a lung X-ray? Or to whom could the affected party turn when ML comes up with a false prediction? The further down we move along the chain that leads towards clinical decision-making, discussed above, the more ethical and legal barriers the ML practitioner/company faces. To mitigate some of these issues, ML systems should be auditable, manufacturers should clearly disclose the characteristics of the database used for training and equip their models with tools that allow reconstructing the reasoning behind the decision taken.

The ML "hype" accompanied by all the recent exaggerated statements in press headlines on its application to healthcare has fuelled a quite controversial discussion, and sometimes human rejection, up to the point

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

that physicians may feel threatened by ML overtaking their jobs in the near future. Moreover, it has been predicted that ML will take the jobs of radiologists and robots will surpass the skills of surgeons. Far from this, ML will not replace physicians, yet medical practitioners who use it will likely substitute those who do not (Langlotz, 2019).

Table 3 organizes the ideas discussed for state interpretation in the form of a SWOT analysis.

Table 3. SWOT analysis – state interpretation

Strengths	**Weaknesses**
• Allow objective and thorough comparison to populations • Allow the integration of complex, heterogenous features • May improve predictive capabilities	• Need well-curated, representative databases for training • Need to extract meaningful, interpretable concepts • Need thorough validation–prospective trials • Need to integrate longitudinal data • Affected by data reliability, representativeness, completeness, and bias • Face ethical/legal barriers and security/regulatory aspects • Ensure transference of knowledge across populations
Opportunities	**Threats**
• Stimulate the man/machine collaboration • Reach diagnosis in a shorter time • Separate ambiguous cases that deserve more attention from clear cases–triaging • Help in the organization of healthcare – diagnosis, risk assessment and urgency assessment	• Harm patients if wrong decisions are taken – high-risk • Disappoint users, especially after all the striking news on ML failures

4. Decision-making (prediction)

Based on the interpretation of the state of the patient, clinicians should take a decision, which either consists in: 1) observing the patient and waiting until an event triggers the need for a decision; 2) collecting more data to improve the odds of taking an informed and appropriate decision; or 3) performing an intervention, followed by monitoring to assess the outcome of the treatment. Machine learning methods can help the clinician to decide which pathway to follow (Funkner, Yakovlev, & Kovalchuk, 2017), in a way that is cost-effective (Morid, Kawamoto, Ault, Dorius, & Abdelrahman, 2017). Making a prediction is perhaps the most difficult stage of the decision-making pipeline, as prescribing treatment implies learning what is the risk associated with each individual.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Several studies have assessed the predictive power of ML. In (Zech et al., 2018), the ML system learns optimal treatment strategies for sepsis in intensive care patients, also providing individualized and clinically interpretable decisions. In another example, supervised reinforcement learning was carried out using recurrent neural networks for recommending treatment for patients admitted to an intensive care unit (Wang, Zhang, He, Tech, & Zha, 2017). An effort towards recommending the ideal invasive intervention for these patients and understanding its outcome using DL has also been reported (Suresh et al., 2017). Its authors were able to provide explanations of the DL decision by exploring the data inputs that maximally activated the CNN outputs. In ophthalmology, a field where ML has performed strongly (Gulshan et al., 2016), DL has demonstrated superior performance in making a referral recommendation compared to experts on a range of sight-threatening retinal diseases (De Fauw et al., 2018). Lastly, DL has also been used to leverage the information contained in electronic health records, achieving high accuracy for tasks such as predicting in-hospital mortality, 30-day unplanned readmission, or prolonged length of stay (Rajkomar et al., 2018).

Clinical status

Several ML models have already been able to predict many important clinical outcomes in very specific applications, from Alzheimer's disease (Bhagwat, Viviano, Voineskos, Chakravarty, & Initiative, 2018) to suicide attempts (Walsh, Ribeiro, & Franklin, 2017). However, the use of ML for prediction and clinical decision-making is still in its infancy, as most models are still incapable of making predictions at the individual level (Kent, Steyerberg, & van Klaveren, 2018; Topol, 2019). A huge effort is still to be done towards integration in a clinical environment, interpretability, and validation, for which prospective randomized clinical trials are needed. In this sense, there is still a long way to go before these tools are ready for implementation in wide routine patient care.

Challenges

Making a prediction is perhaps the most difficult stage of the decision-making process, as prescribing treatment implies learning what will happen and to whom will it happen. The specific challenges concerning clinical decision-making are shared with those of state interpretation and are thus placed in the previous section. Furthermore, many of the challenges discussed in the next section (general challenges) do apply to clinical decision-making.

Table 4 organizes the ideas discussed for decision-making in the form of a SWOT analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Table 4. SWOT analysis – decision-making

Strengths	**Weaknesses**
• May enhance the prediction of clinical outcomes • May enhance the prediction of response to treatment • May improve the recommendation of interventions	• Need well-curated, representative databases for training • Affected by data reliability, representativeness, completeness, and bias • Need to prove clinical benefit • Need to be explainable rather than interpretable • Need to be integrated within clinical systems • Need to prove cost-effectiveness
Opportunities	**Threats**
• Lower cost of healthcare by suggesting cost-effective decisions	• Harm patients if wrong decisions are taken – high-risk • Make decisions for the average patient, not at the individual level

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

General challenges

In the following, we discuss the general challenges that may appear when tackling a clinical problem with ML approaches. These are divided into different sections, depending on whether the challenges are related to the learning itself, the auditability/traceability aspects, the system/infrastructure, or the integration within clinical processes.

Learning

(Non-standardized) data

Patients' medical data are normally kept in many separate systems, which makes it hard to get all the potentially available data of an individual (over a lifespan), let alone being able to make comparisons at a population level. Despite the goldmine of data that every hospital owns, they are underutilized by care providers and clinical researchers, being one of the reasons that electronic health records mostly contain unstructured text. Machine learning systems could be used to infer structured, standardized information from raw data/text, or raw data could be directly used as input to ML solutions specifically designed for complex data integration (Pathak et al., 2013).

Bias and confounding

Let us illustrate how the hype has surpassed the current achievements of ML science with an example of the Watson supercomputer. Despite IBM's initial promises, the recommendations for cancer treatment provided by this AI system have often shown to be erroneous, and physicians at foreign hospitals have reported that its advice is biased toward the methods of care routinely used in the US (Ross & Swetlitz, 2017). A side-effect of this bias is that it amplifies the present gap in health outcomes between the "haves", whose data are used to train ML algorithms, and the "have-nots". Another side effect of bias is that it leaves the minorities well behind the push towards personalized medicine (Topol, 2019). Luckily enough, there are studies that make sure that all minorities are represented in the training data (Bild et al., 2002; Fonseca et al., 2011), but this should be the rule, not the exception.

There have already been examples of ML solutions that inherit human-like biases (Caliskan, Bryson, & Narayanan, 2017), such as the algorithm predicting future criminals biased by race (Flores, Bechtel, & Lowenkamp, 2016), as a consequence of a biased training dataset. Other examples concerning gender bias in medical misdiagnosis exist (Dusenbery, 2018), such as that revealing unequal care after heart attack among sex groups (Wilkinson et al., 2019). This inherited bias occurs because we ask ML solutions to predict which decisions the humans profiled in the training data would have made. Thus, we should not expect the ML method to be fair or impartial or to have the slightest idea about what the clinical goal is. The challenge

is to find the way in which ML overcomes human bias, as this is crucial for successful decision-making applications that do not learn the mistakes that we have committed in the past. A well-known example of bias concerning healthcare is that of risk assessment and prognosis in pneumonia, where ML algorithms classified patients with asthma as being of low-risk and not needing interventions. This erroneous finding derived from the training set, where asthma patients with suspected pneumonia were treated much more aggressively at presentation, thus their risk compared to non-asthma individuals was effectively lowered (Caruana et al., 2015). Despite these hazardous examples of bias, until now, hardly anyone has tried to solve this huge problem. It was not until recently, that scientists from Czech Republic and Germany conducted research to understand the effect of human cognitive bias in ML algorithms, and proposed methods for “debiasing” them (Kliegr, Stěpán Bahník, & Fürnkranz, 2018).

Similar to bias, the learning process can be undermined by confounding, i.e., the finding of a spurious association between the input data and the outcome under study. A simple illustration of this problem could be a DL model that detects pneumonia on chest radiographs, but in reality it learns that certain machines are used, in certain places and ways, on patients who are likely to have pneumonia (Zech et al., 2018). This issue obviously plays against generalizability of the learning model. To deal with this problem, unsupervised learning models, rather than forcing the output to match the ground truth as in supervised applications, let the input data speak for themselves, ordering training examples driven purely by similarity (Hastie, Tibshirani, & Friedman, 2001). If there is a suspicion that confounding effects would still remain, randomization of experiments is highly recommended (Pourhoseingholi, Baghestani, & Vahedi, 2012).

Validation and continuous improvement

Even if an algorithm proves to be well beyond human capacity in prediction tasks, systematic debugging, audit and extensive validation should be mandatory. For ML algorithms to be deployed in hospitals, they must demonstrate improved patient outcomes as well as financial outcomes (Topol, 2019). The core of this validation is multi-centric randomized prospective trials, which are needed to determine how models trained at one site can be best applied to another site. Examples of prospective ML trials assessed in a "real world" clinical environment are scarce–only 6% of 516 surveyed studies performed external validation, according to (Kim, Jang, Kim, Shin, & Park, 2019). In (Lehman et al., 2019), they assessed mammographic breast density using DL and demonstrated good agreement (Cohen’s kappa = 0.85) with radiologists in the prospective validation. In (Steiner et al., 2018) the authors assessed the impact of DL to assist pathologists to review lymph nodes for metastatic breast cancer, and demonstrated the potential of the decision support tool to improve accuracy (sensitivity 91% vs. 83%, P = 0.02) and efficiency–review time per image was shorter with assistance than without it (61 vs. 116 s, P = 0.002). In (Abràmoff, Lavin, Birch, Shah, & Folk, 2018), a ML-based diagnostic system for detection of diabetic retinopathy was prospectively validated. Due to the excellent results achieved, it became the first ML-based system cleared by the FDA for use in clinical practice. In (Long et al., 2017), a ML platform for the diagnosis, risk stratification and treatment suggestions in the context of congenital cataracts was deployed. Last, (Mori et al., 2018) developed and extensively validated a ML model to identify neoplastic polyps requiring resection during colonoscopy in real-time.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Furthermore, one of the greatest benefits of ML models resides in their ability to learn from experience, i.e., to improve their performance as more data become available. However, this might be challenging particularly for neural networks, which have been shown to be prone to issues such as “catastrophic forgetting”– to abruptly forget previously learned information upon learning from new data. Furthermore, re-training on the whole database is time and resource consuming. To solve these problems, federated learning, a novel de-centralized computational architecture where mobile phones run models locally to improve them with a single user‘s data (Konečný et al., 2016), could be helpful. Indeed, this learning strategy has enabled deep neural networks to segment brain tumours without sharing patient data (Sheller, Reina, Edwards, Martin, & Bakas, 2018). Given the evolving nature of ML models, medical device providers are obliged to periodically monitor the performance of their programs, and regulations to evaluate these updates are already being developed (FDA, 2019).

Auditability/traceability

Interpretability vs explainability

Understanding interpretability as the ability to explain or to present in understandable terms to a human (Doshi-Velez & Kim, 2017), providing the user with the ability to interpret the ML’s output is a highly valued characteristic of any learning algorithm, and even more in a strictly-regulated field such as medicine, where lack of interpretability has been described as one of the main limitations hindering the adoption of ML in clinical practice (Cabitza, Rasoini, & Gensini, 2017). Indeed, from the asthma example discussed above (Caruana et al., 2015), it is evident that the non-intelligible use of ML outputs on large datasets can lead to controversial results and therefore direct translation to clinical practice has to be done very cautiously. Unfortunately, many ML implementations available are still opaque and fail to surpass the filter imposed by the European General Data Protection Regulation (GDPR), which compels automated decision-making providers to reveal what is the information and logic involved in each decision (Goodman & Flaxman, 2017).

There is an entire branch of ML research that attempts at enabling the user to access the reasoning path that the algorithm followed to come up with a decision. In (H. Lee et al., 2018), a DL algorithm was used to detect acute intracranial hemorrhages in head CT scans. This system included an attention map and a prediction basis fetched from training data to enhance explainability. In (Rajkomar et al., 2018), the authors revealed which data the model “looked at” for each individual patient, to aid clinicians in determining if a conclusion was reached based on credible facts and to potentially help them on decision-making. Zhang and colleagues (Zhang et al., 2018) recently presented a DL approach for classifying liver lesions on MRI scans that explains the reason for its decisions. In (Lundberg et al., 2018), the authors implemented an explainable ML system to prevent the development of hypoxemia during surgery, using a combination of gradient-boosting machines for prediction and estimates of feature importance for interpretability. Probably one of the methods that has captured most attention in terms of explainability of ML models is LIME, which stands for local interpretable

model-agnostic explanations (Ribeiro, Singh, & Guestrin, 2016). The key intuition behind LIME is that it is much easier to approximate a black-box model by a simple model locally (in the neighbourhood of the prediction we want to explain), as opposed to trying to approximate a model as a whole. To explain an individual prediction, LIME alters the input by changing components that make sense to humans and evaluate how the predictions change. This method allowed to interpret ML-based predictions of individuals at risk of developing hypertension based on cardiorespiratory data (Elshawi, Al-Mallah, & Sakr, 2019).

However, caution is needed with this entire research trend. Explainability is not a synonym of interpretability (Rudin, 2018). The first usually uses black box models, and then draw explanations by saliency maps or estimates of feature importance. Explainable models tend to reach conclusions by fast/intuitive reasoning (see Introduction). These models are appealing, as they are believed to mysteriously uncover unidentified associations in the data as well as to be more accurate, and are “easy” to train. Contrary to black box models, interpretability actually involves a slow/reasoned approach throughout the entire learning path. In this sense, explainability might incur in Type 1 diagnostic error (see Introduction), which happens when the clinical reasoning is driven by intuition–fast and prone to error. On the other hand, interpretability leads to the less common Type 2 error, which happens when the reasoning is analytical, i.e., deductive, deliberate, consistent and scientific (Croskerry, 2009).

Despite all these attempts to enhance interpretability of complex models, this field of research is still in its infancy. A potential solution is generative synthesis (Lundberg et al., 2018), a technology that uses ML to understand neural networks in a fundamental way. It develops a mathematical representation of the studied model and uses it to generate a simplified version of the original neural network that is as accurate but also more compact and faster. Other attempts of mathematically explaining the inner working of neural networks have been undergone (Mallat, 2016). This provides the user with key insights into why and how a network behaves the way it does, thus unravelling the black box enigma (Kumar, Taylor, & Wong, 2017).

For ML models to be applied in clinical decision-making, they cannot merely be interpretable, but they also must be credible. A credible model is an interpretable model that: 1) provides arguments for its predictions that are, at least in part, in-line with domain knowledge; 2) is at least as good as previous standards in terms of predictive performance (J. Wang, Oh, Wang, & Wiens, 2017). For ML models to achieve credibility, it has also been argued that we need to include the human, in this case the medical expert, in the interpretation loop. In practical terms, it would require the data scientist to request from the medical expert a statement of the medical requirements concerning interpretability, a realistic understanding of the interpretability limitations of the ML models, and a description of the decision-making processes at the point of care. Likewise, it would require the medical expert to request from the data scientist guarantees of interpretability adapted to the requirements of the medical problem, compliance with clinical protocols and with system-human interaction workflows (Vellido, 2019).

Together with interpretable models, we should also develop strategies to objectively evaluate the interpretability of a model. The following characteristics may help in the evaluation:

1) Fairness, i.e., we do not want our algorithm to discriminate. There are formal criteria for evaluating fairness (Hardt, Price, & Srebro, 2016).
2) Privacy, i.e., we want our algorithm to protect the privacy of the data it learns from. This also has formal evaluation criteria (Hardt & Talwar, 2009).
3) Reliability, i.e., is our algorithm robust in real-life scenarios? Does it generalize well? Is it vulnerable to adversarial attacks (e.g. ML generated examples that manipulate the decision of another ML system)?
4) Causality, i.e., do the associations learned reflect true causes rather than spurious correlations?
5) Trust, i.e., are the algorithms right for the right reasons, or just by chance?

In (Doshi-Velez & Kim, 2017), the authors further reflect on the concept of interpretability and how to evaluate it for benchmarking purposes. They wonder whether all applications have the same interpretability needs. In the particular case of ML applications in healthcare, we believe that the state interpretation and decision-making blocks should be equipped with the most-advanced interpretability tools, unlike the data acquisition and feature extraction.

However, could this crusade for transparency of learning models be harmful and hamper progress, e.g., due to the disclosure of the trade secret? This should not be the case, as interpretability is not synonymous with transparency. Indeed, (Lipton, 2016) define interpretability as the quality that allows grasping how the models work, and refer to transparency as the answer to performance unknowns, such as whether the model will converge, will produce a unique solution, or whether we understand what each parameter represents. Thus, interpretability should aid post-hoc interpretations, i.e., explain predictions without elucidating the mechanisms. The same applies to human beings, as human decisions allow post-hoc interpretation despite the black box nature of our brains. This notion of informativeness is key to ML applied to healthcare–e.g., by pointing at similar cases, a diagnosis model could provide a hint to a human decision-maker in favour of a diagnostic decision (André, Vercauteren, & Ayache, 2012).

Causal ML rather than predictive ML

Closely linked with the previous point, predictive ML models that are based on correlations of input data and outcomes may not be enough to truly impact the healthcare system. Indeed, predictive ML relying on correlations can be misleading if important causal variables are not considered within the analysis. Should we stop creating predictive algorithms without moving on to finding the root causes of the WHY? Another important aspect for auditability is the ability to interpret the process followed by the algorithm to reach that (diagnostic) decision, i.e., HOW the diagnosis has been made. These two questions are addressed by causal ML, a powerful type of analysis aimed at inferring the mechanisms of the system producing the output data, as if it was learning the transfer function of a filter. In practice, the resulting models would be like maps of how the different variables interact with each other and, once understood, the users can simulate cause and effect of future actions (Pearl & Mackenzie, 2018).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

System-related

Security

Machine learning raises a handful of challenges around data security and privacy, mainly related to the fact that DL models need access to enormous datasets for training purposes. What is the most secure way to transfer large collections of data between different healthcare organizations is still under debate, and stakeholders are no longer underestimating the hazards of a high-profile data leakage. The European GDPR, which recently entered into force, compels to adopt security measures that protect sensible data against hacking and data breaches. Even more harmful is the threat of somebody deliberately hacking a decision-making model to damage people at a large scale.

Information technology resources that adhere to industry standards and regulations are required to warrant the security and privacy of patient data. A potential solution that has been largely discussed is Blockchain, a technology that enables data exchange systems that are cryptographically secured and irrevocable. Blockchain provides a public and immutable log of transactions, cryptographic tools for data integrity and security, and “smart contracts” to regulate data access. The downsides of Blockchain’s technology are that it is slower, maintenance is very costly, and scaling is hard, as data has to be placed in every node of the network rather than in a single place. Even though users are sovereign, which could be an advantage to avoid pernicious companies to access our data, if the users misbehave, there is no easy way to expel them (Song, 2018). Federated learning could potentially be a viable solution to guarantee the security of patient data (see “Validation and continuous improvement subsection”), as this de-centralized model-training paradigm allows to update a learning model without sharing individual information (Silva et al., 2018).

Full anonymization is self-defeating (and maybe impossible to obtain (Rocher, Hendrickx, & de Montjoye, 2019)) as it causes failure when explanations are needed. Machine learning systems that deal with/implement pseudo-anonymization are key. A highly desirable feature is that they are also auditable.

Regulatory

The use of ML for clinical decision-making will unavoidably bring legal challenges regarding medical negligence derived from learning failures. When malpractice cases involving medical ML applications arise, the legal system needs to provide clear guidance on what entity holds liability, however regulations in this arena still lag behind. Furthermore, ML in healthcare poses a unique challenge to regulatory agencies because the models can quickly evolve as more data and user feedback are collected, and it is not clear how the updates should be evaluated (FDA, 2019). Policymakers should guide and generate specific criteria for the process of demonstrating non-inferiority of algorithms compared to existing standards, specially emphasizing on the validation process, and quality of the training and validation data (Yu, Beam, & Kohane, 2018). We need to establish adequate laws that ensure that algorithms are used properly and for people’s

welfare. In summary, for ML technology to be adopted by healthcare systems within the next years many legal aspects still need to be addressed, and decision- and policy-makers should join efforts towards this end.

Integration (man/machine coexistence)

The vision of ML tools replacing humans in clinical medicine may be a matter of concern for some healthcare professionals, especially radiologists (Obermeyer & Emanuel, 2016). To the relief of many, this end-to-end solution scenario is highly unlikely in the near to mid-term future. A more realistic scenario is that of learning algorithms targeting repetitive sub-tasks, to assist physicians to reach a more informed decision or to help them become more efficient. On top of the formidable obstacles and pitfalls of current ML solutions discussed above (Topol, 2019), clinicians will still be needed to interact with the patients by taking targeted patient histories and performing physical examinations, navigate diagnostic procedures, integrate solutions proposed by ML systems and adapt them according to the changing stages of disease or patient's preferences, inform the patient's family about therapy options, or console them if the disease stage is very advanced.

Accordingly, we should quickly abandon the idea of human-machine competition and rather think of a cooperation paradigm, where to exploit collaboration and synergies between human intelligence and ML. Indeed, ML and humans possess complementary skills: ML techniques stand out at computation and pattern recognition on massive amounts of data (Type 1 reasoning), whereas people are far better at understanding the context, abstracting knowledge from their experience, and transferring it across domains (Type 2 reasoning). The emphasis should now be on human-in-the-loop approaches that enable users to interact effectively with ML models to make better decisions based on large datasets, without requiring in-depth technical knowledge about the model inner-workings. However, understanding where automated systems can be used, and which level of automation would be appropriate within clinical procedures is crucial to avoid potentially preventable errors attributed to automation bias (Cummings, 2004). Examples of human-machine collaboration already exist. In diabetic retinopathy diagnosis (Sayres et al., 2018), model assistance increased the accuracy of retina specialists above that of the unassisted reader or model alone. Others examples have demonstrated the synergy of man/machine by improving diagnosis of knee injuries using MRI (Bien et al., 2018).

In light of this, the current clinical workflow could be rethought: a diagnosis is proposed by the ML system, the human operator verifies the data on which the conclusions are drawn, informs the system of potential measurement errors or confounders, and finally accepts or rejects the diagnosis. Thus, the human operator preserves the overall control, while machines perform measurements and integrate and compare data at request (D'hooge & Fraser, 2018). Ultimately, this human-machine symbiosis will be beneficial to release physicians from low-level tasks such as measurements, data preparation, standardization, to give them more time on higher-level tasks such as patient care and clinical decision-making (Sengupta & Adjeroh, 2018).

ML applied to real clinical data

In human decision-making, a clinician would explore all available data and use experience to compare them to patients they have seen before or were trained to recognize. Once an individual is put into context with regards to expected normality and typical cases, previous knowledge on treatment effect is used to manage this certain patient. This 'eminence-based' approach is only within reach of a few experienced clinicians. For simplification and standardization, many professional organizations provide diagnostic guidelines based on data from large cohorts or clinical trials (Brignole et al., 2013; Epstein et al., 2008; Ponikowski et al., 2016). Although these recommendations have proven to standardize medical care in a better way, the current process of formulating guidelines is not guaranteed to make best use of the full complex original data available. In this sense, the use of ML seems amply justified.

Still, a significant amount of papers on ML for decision-making deal with data collected following strict input criteria and well-defined protocols used in randomized clinical trials (Ambale-Venkatesh et al., 2017; Kalscheur et al., 2018). However, clinical practice is often much more complex and varied as compared to clinical trials. In clinical reality data are incomplete, and ML techniques need to be able to deal with incompleteness, either by performing imputation or by adopting formulations that explicitly take into account that the data can be incomplete. Furthermore, patients often lay outside the narrow selection criteria of trials (including co-morbidities, ethnicity, gender, age, lifestyle, etc.), may have been treated before the investigation using different protocols, may present at a different stage of disease, and most importantly, may undergo different decision pathways and be seen by different individuals, each with their own decision-making process, during the study. On top of this, obtaining a hard outcome to train an algorithm is often difficult, for example to register death, the study would need to be carried on until everybody dies, which is unfeasible both for time and economic constraints. Even if registered, often the outcome is scarce (e.g., number of deaths caused by a heart attack), and when appearing, the reason for experiencing it is different for every patient– as the decision pathway has likely been different (Oladapo et al., 2015).

All these aspects make it extremely challenging to associate input descriptors to outcomes using supervised predictive ML/DL techniques that fail to understand/capture the context from which data have been drawn. Indeed, their blind application on large datasets likely yields unwanted results, which might even be dangerous for the patient. In such unfavourable scenarios, a more promising approach could be based on data dimensionality reduction, an unsupervised, label-agnostic approach where all data are used to position individuals according to their similarity regarding the input descriptors, and this similarity combined with previous knowledge can be used to infer diagnosis or treatment response.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Conclusion

For ML to effectively interpret clinical data, it logically needs to follow, or be closely integrated in, the clinical decision pathway used by physicians in routine practice. Furthermore, ML needs to learn the context of the problem and quality and meaning of data, which makes it a non-trivial task. Therefore, we should evolve from a weak –where the machine learns a single, repetitive task– towards a strong ML paradigm. Nevertheless, for better or for worse, the achievement of the level of intelligence to cover the whole decision-making process, is unlikely in the short to medium term. Consequently, the foreseeable application of ML in healthcare consists of different technologies/algorithms capable of performing equal or better than humans in specific/well-defined tasks, so that it increases consistency and helps improving decision-making, with the ultimate goal of supporting physicians in the management of the patients.

The above can only be achieved by multidisciplinary teams making a joint, closely integrated effort. We consider that the upcoming policies for ML research in healthcare should address the challenges described in the previous section. Specifically, much work is needed on formally analysing which are the intrinsic processes being used during clinical decision-making, to further explore and identify the necessities where ML can help. On the algorithms themselves, a bulk of research should be dedicated to dealing with longitudinal data, how to best describe a patient, and how to relate the learning with pathophysiology, i.e., how can we marry previous clinical knowledge with the algorithmic conclusions. Data integration and what is the best approach for dealing with incomplete data and outliers should be also surveyed. We need to augment current validation techniques with additional components to quantify generalization performance and develop uncertainty quantification methods to establish trust in the (predictive) models. We need as well to develop metrics to determine whether or not ML has a positive impact on patients' outcomes, instead of letting us be driven by the hype of using ML everywhere. Finally, we need to explore the practical considerations that will affect adoption of the ML technology, such as how ML software should be integrated with PACS, or how it would be paid for by facilities. For these, a clear demonstration of the cost-effectiveness of ML technology in healthcare systems is needed.

Beyond all the aforementioned obstacles and pitfalls, the new ML era is called to revolutionize healthcare as we know it. ML applications will surely help reducing the stress of the healthcare personnel and improving the quality of healthcare systems worldwide, and will contribute to optimal clinical decision-making, if considered with care. We anticipate exciting times for medicine ahead.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 November 2019 doi:10.20944/preprints201911.0278.v1

Bibliography

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Npj Digital Medicine*, *1*(1), 39. https://doi.org/10.1038/s41746-018-0040-6

Ambale-Venkatesh, B., Yang, X., Wu, C. O., Liu, K., Gregory Hundley, W., McClelland, R., … Lima, J. A. C. (2017). Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. *Circulation Research*, *121*(9), 1092–1101. https://doi.org/10.1161/CIRCRESAHA.117.311312

André, B., Vercauteren, T., & Ayache, N. (2012). Content-based retrieval in endomicroscopy: toward an efficient smart atlas for clinical diagnosis. In *MICCAI – MCBR-CDS 2011* (pp. 12–23). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-28460-1_2

Annarumma, M., Withey, S. J., Bakewell, R. J., Pesce, E., Goh, V., & Montana, G. (2019). Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology*, 180921. https://doi.org/10.1148/radiol.2018180921

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2019). VoxelMorph: a Llearning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, *38*(8), 1788–1800. https://doi.org/10.1109/TMI.2019.2897538

Beecy, A. N., Chang, Q., Anchouche, K., Baskaran, L., Elmore, K., Kolli, K., … Min, J. K. (2018). A novel deep learning approach for automated diagnosis of acute ischemic infarction on computed tomography. *JACC: Cardiovascular Imaging*, *11*(11), 1723–1725. https://doi.org/10.1016/J.JCMG.2018.03.012

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, *1*(1), 20–23. https://doi.org/10.1038/s42256-018-0004-1

Bellver, M., Maninis, K.-K., Pont-Tuset, J., Giro-i-Nieto, X., Torres, J., & Van Gool, L. (2017). Detection-aided liver lesion segmentation using deep learning. *ArXiv*. Retrieved from http://arxiv.org/abs/1711.11069

Benavidez, O. J., Gauvreau, K., & Geva, T. (2014). Diagnostic errors in congenital echocardiography: importance of study conditions. *Journal of the American Society of Echocardiography*, *27*(6), 616–623. https://doi.org/10.1016/j.echo.2014.03.001

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., … Jodoin, P.-M. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, *37*(11), 2514–2525. https://doi.org/10.1109/TMI.2018.2837502

Bhagwat, N., Viviano, J. D., Voineskos, A. N., Chakravarty, M. M., & Initiative, A. D. N. (2018). Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLOS Computational Biology*, *14*(9), e1006376. https://doi.org/10.1371/journal.pcbi.1006376

Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., … Lungren, M. P. (2018). Deep-learning-

assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLOS Medicine*, *15*(11), e1002699. https://doi.org/10.1371/journal.pmed.1002699

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., … Tracy, R. P. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology*, *156*(9), 871–881. https://doi.org/10.1093/aje/kwf113

Bizzego, A., Bussola, N., Salvalai, D., Chierici, M., Maggio, V., Jurman, G., & Furlanello, C. (2019). Integrating deep and radiomics features in cancer bioimaging. In *IEEE Conference on Computational intelligence in Bioinformatics and Computational Biology*. Cold Spring Harbor Laboratory. https://doi.org/10.1109/CIBCB.2019.8791473

Brignole, M., Auricchio, A., Baron-Esquivias, G., Bordachar, P., Boriani, G., Breithardt, O. A., … Vardas, P. E. (2013). 2013 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy. *European Heart Journal*, *15*(8), 1070–1118. https://doi.org/10.1093/europace/eut206

Brosch, T., Tang, L. Y. W., Yoo, Y., Li, D. K. B., Traboulsee, A., & Tam, R. (2016). Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Transactions on Medical Imaging*, *35*(5), 1229–1239. https://doi.org/10.1109/TMI.2016.2528821

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA*, *318*(6), 517. https://doi.org/10.1001/jama.2017.7797

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15* (pp. 1721–1730). New York, New York, USA: ACM Press. https://doi.org/10.1145/2783258.2788613

Chen, L., Bentley, P., & Rueckert, D. (2017). Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *NeuroImage: Clinical*, *15*, 633–643. https://doi.org/10.1016/J.NICL.2017.06.016

Cikes, M., Sanchez-Martinez, S., Claggett, B., Duchateau, N., Piella, G., Butakoff, C., … Bijnens, B. (2019). Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *European Journal of Heart Failure*, *21*(1), 74–85. https://doi.org/10.1002/ejhf.1333

Cikes, M., & Solomon, S. D. (2016). Beyond ejection fraction: An integrative approach for assessment of cardiac structure and function in heart failure. *European Heart Journal*, *37*(21), 1642–1650. https://doi.org/10.1093/eurheartj/ehv510

Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Corinne Abrams. (2019). Google's effort to prevent blindness shows AI challenges. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/googles-effort-to-prevent-blindness-hits-roadblock-11548504004

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine*, *84*(8), 1022–1028. https://doi.org/10.1097/ACM.0b013e3181ace703

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *AIAA 3RD INTELLIGENT SYSTEMS CONFERENCE* (pp. 2004--6313). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.2634

D'hooge, J., & Fraser, A. G. (2018). Learning about machine learning to create a self-driving echocardiographic laboratory. *Circulation*, *138*(16), 1636–1638. https://doi.org/10.1161/CIRCULATIONAHA.118.037094

Dalca, A. V., Balakrishnan, G., Guttag, J., & Sabuncu, M. R. (2019). Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, *57*, 226–236. https://doi.org/https://doi.org/10.1016/j.media.2019.07.006

Danelakis, A., Theoharis, T., & Verganelakis, D. A. (2018). Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, *70*, 83–100. https://doi.org/10.1016/J.COMPMEDIMAG.2018.10.002

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, *24*(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

Desai, A. S., Jhund, P. S., Yancy, C., Lopatin, M., Stevenson, L., Marco, T. De, … Investigators, R. L.-H. T. (2016). After TOPCAT: What to do now in Heart Failure with Preserved Ejection Fraction. *European Heart Journal*, *47*(9), 1510–1518. https://doi.org/10.1093/eurheartj/ehw114

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., … Franc, B. L. (2018). A deep learning model to predict a diagnosis of Alzheimer disease by using 18 F-FDG PET of the brain. *Radiology*, 180958. https://doi.org/10.1148/radiol.2018180958

Doshi-Velez, F., Ge, Y., & Kohane, I. (2014). Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, *133*(1), e54-63. https://doi.org/10.1542/peds.2013-0819

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv*. Retrieved from http://arxiv.org/abs/1702.08608

Dusenbery, M. (2018). *Doing harm: the truth about how bad medicine and lazy science leave women dismissed, misdiagnosed, and sick*. HarperOne.

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-0874-0

Epstein, A. E., Dimarco, J. P., Ellenbogen, K. A., Estes, N. A. M., Freedman, R. A., Gettes, L. S., … Yancy, C. W. (2008). ACC/AHA/HRS 2008 Guidelines of cardiac rhythm abnormalities. A report of the American College of Cardiology/American Heart Association task force on practice guidelines (writing committee to revise the ACC/AHA/NASPE 2002 Guideline update for implantation. *Circulation*, *117*, 350–408. https://doi.org/10.1161/CIRCUALTIONAHA.108.189742

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

https://doi.org/10.1038/nature21056

FDA. (2019). *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)-discussion paper and request for feedback*. Retrieved from https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm514737.pdf.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: a rejoinder to “machine bias: there’s software used across the country to predict future criminals. And it’s biased against blacks.” *Federal Probation Journal*, *80*(2), 38–46. Retrieved from https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder

Fonseca, C. G., Backhaus, M., Bluemke, D. A., Britten, R. D., Chung, J. Do, Cowan, B. R., … Young, A. A. (2011). The Cardiac Atlas Project—an imaging database for computational modeling and statistical atlases of the heart. *Bioinformatics*, *27*(16), 2288–2295. https://doi.org/10.1093/bioinformatics/btr360

Funkner, A. A., Yakovlev, A. N., & Kovalchuk, S. V. (2017). Data-driven modeling of clinical pathways using electronic health records. *Procedia Computer Science*, *121*, 835–842. https://doi.org/10.1016/j.procs.2017.11.108

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association : JAMIA*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. Retrieved from https://www.deeplearningbook.org/

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation.” *AI Magazine*, *38*(3). https://doi.org/10.1609/aimag.v38i3.2741

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402. https://doi.org/10.1001/jama.2016.17216

Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M., & Notar, M. (2018). An application of machine learning to haematological diagnosis. *Scientific Reports*, *8*(1), 411. https://doi.org/10.1038/s41598-017-18564-8

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., … Zalaudek, I. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842. https://doi.org/10.1093/annonc/mdy166

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS*. Retrieved from http://arxiv.org/abs/1610.02413

Hardt, M., & Talwar, K. (2009). On the geometry of differential privacy. *ArXiv*. Retrieved from http://arxiv.org/abs/0907.3754

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning (Vol.1)*. Springer, Berlin: Springer series in statistics.

Housden, J., Wang, S., Noh, Y., Singh, D., Singh, A., Back, J., … Rhode, K. (2017). Control strategy for a new extra-corporeal robotic ultrasound system. In *MPEC 2017*.

Janowczyk, A., Basavanhally, A., & Madabhushi, A. (2017). Stain normalization using sparse autoEncoders (StaNoSA): application to digital pathology. *Computerized Medical Imaging and Graphics*, *57*, 50–61. https://doi.org/10.1016/J.COMPMEDIMAG.2016.05.003

Kalscheur, M. M., Kipp, R. T., Tattersall, M. C., Mei, C., Buhr, K. A., DeMets, D. L., … Page, C. D. (2018). Machine Learning Algorithm Predicts Cardiac Resynchronization Therapy Outcomes. *Circulation: Arrhythmia and Electrophysiology*, *11*(1). https://doi.org/10.1161/CIRCEP.117.005499

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., … Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78. https://doi.org/10.1016/J.MEDIA.2016.10.004

Kang, E., Min, J., & Ye, J. C. (2017). A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. *Medical Physics*, *44*(10), e360–e375. https://doi.org/10.1002/mp.12344

Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., & Mukhopadhyay, A. (2018). GANs for medical image analysis. *ArXiv*. Retrieved from http://arxiv.org/abs/1809.06222

Kent, D. M., Steyerberg, E., & van Klaveren, D. (2018). Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*, *363*. https://doi.org/10.1136/bmj.k4245

Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y., & Park, S. H. (2019). Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean Journal of Radiology*, *20*(3), 405. https://doi.org/10.3348/kjr.2019.0025

Kliegr, T., Stěpán Bahník, ˇ, & Fürnkranz, J. (2018). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *ArXiv*. Retrieved from https://arxiv.org/pdf/1804.02969.pdf

Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: strategies for improving communication efficiency. *ArXiv*. Retrieved from https://ai.google/research/pubs/pub45648

Krebs, J., Delingette, H., Mailhe, B., Ayache, N., & Mansi, T. (2019). Learning a probabilistic model for diffeomorphic registration. *IEEE Transactions on Medical Imaging*, *38*(9), 2165–2176. https://doi.org/10.1109/TMI.2019.2897112

Kumar, D., Taylor, G. W., & Wong, A. (2017). Discovery radiomics with CLEAR-DR: interpretable computer aided diagnosis of diabetic retinopathy. *IEEE Access*, *7*, 25891–25896. https://doi.org/10.1109/ACCESS.2019.2893635

Kutyifa, V., Kloppe, A., Zareba, W., Solomon, S. D., McNitt, S., Polonsky, S., … Goldenberg, I. (2013). The influence of left ventricular ejection fraction on the effectiveness of cardiac resynchronization therapy: MADIT-CRT (multicenter automatic defibrillator implantation trial with cardiac resynchronization therapy). *Journal of the American College of Cardiology*, *61*(9), 936–944. https://doi.org/10.1016/J.JACC.2012.11.051

Lam, C. S. P., & Solomon, S. D. (2014). The middle child in heart failure: heart failure with mid-range

ejection fraction (40-50%). *European Journal of Heart Failure*, *16*(10), 1049–1055. https://doi.org/10.1002/ejhf.159

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R. G. P. M., Granton, P., … Aerts, H. J. W. L. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, *48*(4), 441–446. https://doi.org/10.1016/J.EJCA.2011.11.036

Langlotz, C. P. (2019). Will Artificial Intelligence Replace Radiologists? *Radiology: Artificial Intelligence*, *1*(3), e190058. https://doi.org/10.1148/ryai.2019190058

Lao, J., Chen, Y., Li, Z.-C., Li, Q., Zhang, J., Liu, J., & Zhai, G. (2017). A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific Reports*, *7*(1), 10353. https://doi.org/10.1038/s41598-017-10649-8

Lee, G., Kang, B., Nho, K., Sohn, K.-A., & Kim, D. (2019). MildInt: deep learning-based multimodal longitudinal data integration framework. *Frontiers in Genetics*, *10*, 617. https://doi.org/10.3389/fgene.2019.00617

Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S. H., Guerrier, C. E., … Do, S. (2018). An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering*, 1. https://doi.org/10.1038/s41551-018-0324-9

Lehman, C. D., Yala, A., Schuster, T., Dontchos, B., Bahl, M., Swanson, K., & Barzilay, R. (2019). Mammographic breast density assessment using deep learning: clinical implementation. *Radiology*, *290*(1), 52–58. https://doi.org/10.1148/radiol.2018180694

Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D., & Ji, S. (2014). Deep learning based imaging data completion for improved brain disease diagnosis. In *MICCAI 2014* (pp. 305–312). Springer, Cham. https://doi.org/10.1007/978-3-319-10443-0_39

Lipton, Z. C. (2016). The mythos of model interpretability. *ArXiv*. Retrieved from http://arxiv.org/abs/1606.03490

Litjens, G., Ciompi, F., Wolterink, J. M., de Vos, B. D., Leiner, T., Teuwen, J., & Išgum, I. (2019). State-of-the-Art Deep Learning in Cardiovascular Image Analysis. *JACC: Cardiovascular Imaging*, *12*(8), 1549–1565. https://doi.org/10.1016/j.jcmg.2019.06.009

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., … Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, *1*(6), e271–e297. https://doi.org/10.1016/s2589-7500(19)30123-2

Long, E., Lin, H., Liu, Z., Wu, X., Wang, L., Jiang, J., … Liu, Y. (2017). An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nature Biomedical Engineering*, *1*(2), 0024. https://doi.org/10.1038/s41551-016-0024

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., … Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, *2*(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0

Madani, A., Arnaout, R., Mofrad, M., & Arnaout, R. (2018). Fast and accurate view classification of echocardiograms using deep learning. *Npj Digital Medicine*, *1*(1), 6. https://doi.org/10.1038/s41746-

017-0013-1

Madani, A., Ong, J. R., Tibrewal, A., & Mofrad, M. R. K. (2018). Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *Npj Digital Medicine*, *1*(1), 59. https://doi.org/10.1038/s41746-018-0065-x

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150203. https://doi.org/10.1098/rsta.2015.0203

Menking, C., & Pien, H. (2018). *WO2018220089A1*. Retrieved from https://patents.google.com/patent/WO2018220089A1/en?q=philips&q=medical+imaging&q=deep+learning&oq=philips+medical+imaging+deep+learning

Miao, S., Wang, Z. J., & Liao, R. (2016). A CNN regression approach for real-time 2D/3D registration. *IEEE Transactions on Medical Imaging*, *35*(5), 1352–1363. https://doi.org/10.1109/TMI.2016.2521800

Mincholé, A., Camps, J., Lyon, A., & Rodríguez, B. (2019). Machine learning in the electrocardiogram. *Journal of Electrocardiology*. https://doi.org/10.1016/J.JELECTROCARD.2019.08.008

Mo, J.-H., & Lu, X.-M. (2005). *US7878977B2*. Retrieved from https://patents.google.com/patent/US7878977B2/en

Moeskops, P., Wolterink, J. M., van der Velden, B. H. M., Gilhuijs, K. G. A., Leiner, T., Viergever, M. A., & Išgum, I. (2016). Deep learning for multi-task medical image segmentation in multiple modalities. *ArXiv*, 478–486. https://doi.org/10.1007/978-3-319-46723-8_55

Mori, Y., Kudo, S., Misawa, M., Saito, Y., Ikematsu, H., Hotta, K., … Mori, K. (2018). Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy. *Annals of Internal Medicine*, *169*(6), 357. https://doi.org/10.7326/M18-0249

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2017). Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. *AMIA Symposium - Proceedings.*, *2017*, 1312–1321. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/29854200

Narula, S., Shameer, K., Salem Omar, A. M., Dudley, J. T., & Sengupta, P. P. (2016). Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. *Journal of the American College of Cardiology*, *68*(21), 2287–2295. https://doi.org/10.1016/j.jacc.2016.08.062

National Academies of Sciences, Engineering, and M. (2015). *Improving diagnosis in health care*. (E. P. Balogh, B. T. Miller, & J. R. Ball, Eds.). Washington, D.C.: National Academies Press. https://doi.org/10.17226/21794

Nie, D., Cao, X., Gao, Y., Wang, L., & Shen, D. (2016). Estimating CT image from MRI data using 3d fully convolutional networks. In *International Workshop on Deep Learning in Medical Image Analysis* (pp. 170–178). Springer, Cham. https://doi.org/10.1007/978-3-319-46976-8_18

Nogueira, M., Piella, G., Sanchez-Martinez, S., Langet, H., Saloux, E., Bijnens, B., & De Craene, M. (2017). Characterizing patterns of response during mild stress-testing in continuous echocardiography recordings using a multiview dimensionality reduction technique. In *Functional Imaging and Modelling of the Heart – conference proceedings* (Vol. 10263 LNCS). https://doi.org/10.1007/978-3-319-59448-4_48

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future – big data, machine learning, and clinical medicine. *New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1609300

Oktay, O., Bai, W., Lee, M., Guerrero, R., Kamnitsas, K., Caballero, J., … Rueckert, D. (2016). Multi-input cardiac image super-resolution using convolutional neural networks. In *MICCAI 2016* (pp. 246–254). Springer, Cham. https://doi.org/10.1007/978-3-319-46726-9_29

Oladapo, O. T., Souza, J. P., Bohren, M. A., Tunçalp, Ö., Vogel, J. P., Fawole, B., … Gülmezoglu, A. M. (2015). WHO Better Outcomes in Labour Difficulty (BOLD) project: innovating to improve quality of care around the time of childbirth. *Reproductive Health*, *12*, 48. https://doi.org/10.1186/s12978-015-0027-6

Østvik, A., Smistad, E., Aase, S. A., Haugen, B. O., & Lovstakken, L. (2019). Real-time standard view classification in transthoracic echocardiography using convolutional neural networks. *Ultrasound in Medicine & Biology*, *45*(2), 374–384. https://doi.org/10.1016/J.ULTRASMEDBIO.2018.07.024

Østvik, A., Smistad, E., Espeland, T., Berg, E. A. R., & Lovstakken, L. (2018). Automatic myocardial strain imaging in echocardiography using deep learning. In D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, … A. Madabhushi (Eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 309–316). Cham: Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-030-00889-5_35

Pathak, J., Bailey, K. R., Beebe, C. E., Bethard, S., Carrell, D. S., Chen, P. J., … Chute, C. G. (2013). Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association*, *20*(e2), e341–e348. https://doi.org/10.1136/amiajnl-2013-001939

Pearl, J., & Mackenzie, D. (2018). *The book of why : the new science of cause and effect* (1st ed.). Basic Books, Inc. Retrieved from https://www.basicbooks.com/titles/judea-pearl/the-book-of-why/9780465097609/

Pellikka, P. A., She, L., Holly, T. A., Lin, G., Varadarajan, P., Pai, R. G., … Oh, J. K. (2018). Variability in ejection fraction measured by echocardiography, gated single-photon emission computed tomography, and cardiac magnetic resonance in patients with coronary artery disease and left ventricular dysfunction. *JAMA Network Open*, *1*(4), e181456. https://doi.org/10.1001/jamanetworkopen.2018.1456

Ponikowski, P., Voors, A. A., Anker, S. D., Bueno, H., González-Juanatey, J. R., Harjola, V.-P., … Meer, P. van der. (2016). 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal*, *37*(27), 2129–2200. https://doi.org/10.1093/eurheartj/ehw128

Pourhoseingholi, M. A., Baghestani, A. R., & Vahedi, M. (2012). How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from Bed to Bench*, *5*(2), 79–83. https://doi.org/https://doi.org/10.22037/ghfbb.v5i2.246

Qin, C., Schlemper, J., Caballero, J., Price, A. N., Hajnal, J. V., & Rueckert, D. (2019). Convolutional recurrent neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, *38*(1), 280–290. https://doi.org/10.1109/TMI.2018.2863670

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., … Dean, J. (2018). Scalable and

accurate deep learning with electronic health records. *Npj Digital Medicine*, *1*(1), 18. https://doi.org/10.1038/s41746-018-0029-1

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": explaining the predictions of any classifier. *ArXiv*. Retrieved from http://arxiv.org/abs/1602.04938

Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*(1), 3069. https://doi.org/10.1038/s41467-019-10933-3

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In *MICCAI 2015* (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Ross, C., & Swetlitz, I. (2017). IBM pitched Watson as a revolution in cancer care. It's nowhere close. *Stat News*. Retrieved from https://www.statnews.com/2017/09/05/watson-ibm-cancer/

Rothberg, J. M., Fife, K. G., Ralston, T. S., Charvat, G. L., & Sanchez, N. J. (2014). *US20140288428A1*. Retrieved from https://patents.google.com/patent/US20140288428A1/en?q=ultrasound-on-chip&oq=ultrasound-on-chip

Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *ArXiv*. Retrieved from http://arxiv.org/abs/1811.10154

Ruijsink, B., Puyol-Antón, E., Oksuz, I., Sinclair, M., Bai, W., Schnabel, J. A., … King, A. P. (2019). Fully automated, quality-controlled cardiac analysis from CMR: validation and large-scale application to characterize cardiac function. *JACC: Cardiovascular Imaging*. https://doi.org/10.1016/J.JCMG.2019.05.030

Sanchez-Martinez, S., Duchateau, N., Erdei, T., Kunszt, G., Aakhus, S., Degiovanni, A., … Bijnens, B. H. (2018). Machine learning analysis of left ventricular function to characterize heart failure with preserved ejection fraction. *Circulation. Cardiovascular Imaging*, *11*(4), e007138. https://doi.org/10.1161/CIRCIMAGING.117.007138

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., … Webster, D. R. (2018). Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*. https://doi.org/10.1016/J.OPHTHA.2018.11.016

Sengupta, P. P., & Adjeroh, D. A. (2018). Will artificial intelligence replace the human echocardiographer? *Circulation*, *138*(16), 1639–1642. https://doi.org/10.1161/CIRCULATIONAHA.118.037095

Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2018). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop* (Vol. 11383 LNCS, pp. 92–104). Springer Verlag. https://doi.org/10.1007/978-3-030-11723-8_9

Shin, H.-C., Le Lu, Kim, L., Seff, A., Yao, J., & Summers, R. M. (2015). Interleaved text/image Deep Mining on a large-scale radiology database. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1090–1099). IEEE. https://doi.org/10.1109/CVPR.2015.7298712

Silva, S., Gutman, B., Romero, E., Thompson, P. M., Altmann, A., & Lorenzi, M. (2018). Federated learning in distributed medical databases: meta-analysis of large-scale subcortical brain data. *ArXiv*. Retrieved from http://arxiv.org/abs/1810.08553

Smuha, N. (2019). A definition of artificial intelligence: main capabilities and scientific disciplines. *European*

Commission Reports, 7. Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341

Song, J. (2018, May). Why blockchain is hard. *Medium*. Retrieved from https://medium.com/@jimmysong/why-blockchain-is-hard-60416ea4c5c

Steiner, D. F., Macdonald, R., Liu, Y., Truszkowski, P., Hipp, J. D., Gammage, C., … Stumpe, M. C. (2018). Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American Journal of Surgical Pathology*, *42*(12), 1636–1646. https://doi.org/10.1097/pas.0000000000001151

Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P., & Ghassemi, M. (2017). Clinical intervention prediction and understanding using deep networks. *ArXiv*. Retrieved from http://arxiv.org/abs/1705.08498

Tom, F., & Sheet, D. (2017). Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning. *ArXiv*. Retrieved from http://www.cvc.uab.es/IVUSchallenge2011/dataset.html

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Triposkiadis, F., Butler, J., Abboud, F. M., Armstrong, P. W., Adamopoulos, S., Atherton, J. J., … De Keulenaer, G. W. (2019). The continuous heart failure spectrum: moving beyond an ejection fraction classification. *European Heart Journal*. https://doi.org/10.1093/eurheartj/ehz158

Tschandl, P., Rosendahl, C., Akay, B. N., Argenziano, G., Blum, A., Braun, R. P., … Kittler, H. (2019). Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatology*, *155*(1), 58. https://doi.org/10.1001/jamadermatol.2018.4378

USAID, C. for I. and I., Rockefeller, F., & Gates, F. (2019). *Artificial intelligence in global health: defining a collective path forward*. Retrieved from https://www.usaid.gov/sites/default/files/documents/1864/AI-in-Global-Health_webFinal_508.pdf

Vedula, S., Senouf, O., Bronstein, A. M., Michailovich, O. V, & Zibulevsky, M. (2017). Towards CT-quality ultrasound imaging using deep learning. *ArXiv*. Retrieved from https://arxiv.org/pdf/1710.06304v1.pdf

Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 1–15. https://doi.org/10.1007/s00521-019-04051-w

Vesal, S., Ravikumar, N., & Maier, A. (2018). SkinNet: a deep learning framework for skin lesion segmentation. *ArXiv*. Retrieved from http://arxiv.org/abs/1806.09522

Voelker, R. (2018). Diagnosing fractures with AI. *JAMA*, *320*(1), 23. https://doi.org/10.1001/jama.2018.8565

Wallentin, L., Gale, C. P., Maggioni, A., Bardinet, I., & Casadei, B. (2019). EuroHeart: European unified registries on heart care evaluation and randomized trials. *European Heart Journal*, *40*(33), 2745–2749. https://doi.org/10.1093/eurheartj/ehz599

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *5*(3), 457–469. https://doi.org/10.1177/2167702617691560

Wang, J., Oh, J., Wang, H., & Wiens, J. (2017). Learning credible models. *ArXiv*. Retrieved from

http://arxiv.org/abs/1711.03190

Wang, L., Zhang, W., He, X., Tech, G., & Zha, H. (2017). Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. *ArXiv*, *17*. Retrieved from https://doi.org/10.1145/nnnnnnn.nnnnnnn

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., … Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 1–4. https://doi.org/10.1038/s41591-019-0548-6

Wilkinson, C., Bebb, O., Dondo, T. B., Munyombwe, T., Casadei, B., Clarke, S., … Gale, C. P. (2019). Sex differences in quality indicator attainment for myocardial infarction: a nationwide cohort study. *Heart (British Cardiac Society)*, *105*(7), 516–523. https://doi.org/10.1136/heartjnl-2018-313959

Wu, L., Cheng, J.-Z., Li, S., Lei, B., Wang, T., & Ni, D. (2017). FUIQA: fetal ultrasound image quality assessment with deep convolutional networks. *IEEE Transactions on Cybernetics*, *47*(5), 1336–1349. https://doi.org/10.1109/TCYB.2017.2671898

Xue, W., Brahm, G., Pandey, S., Leung, S., & Li, S. (2018). Full left ventricle quantification via deep multitask relationships learning. *Medical Image Analysis*, *43*, 54–65. https://doi.org/10.1016/J.MEDIA.2017.09.005

Yang, W., Chen, Y., Liu, Y., Zhong, L., Qin, G., Lu, Z., … Chen, W. (2017). Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. *Medical Image Analysis*, *35*, 421–433. https://doi.org/10.1016/J.MEDIA.2016.08.004

Yoon, Y. H., Khan, S., Huh, J., & Ye, J. C. (2019). Efficient B-Mode Ultrasound Image Reconstruction From Sub-Sampled RF Data Using Deep Learning. *IEEE Transactions on Medical Imaging*, *38*(2), 325–336. https://doi.org/10.1109/TMI.2018.2864821

Yu, K., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(October), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, *15*(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683

Zhang, F., Yang, J., Nezami, N., Laage-gaupp, F., Chapiro, J., De Lin, M., & Duncan, J. (2018). Liver tissue classification using an auto-context-based deep neural network with a multi-phase training framework. In *Patch-MI 2018* (pp. 59–66). https://doi.org/10.1007/978-3-030-00500-9_7

Zhang, J., Gajjala, S., Agrawal, P., Tison, G. H., Hallock, L. A., Beussink-Nelson, L., … Deo, R. C. (2018). Fully automated echocardiogram interpretation in clinical practice. *Circulation*, *138*(16), 1623–1635. https://doi.org/10.1161/CIRCULATIONAHA.118.034338

Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., & Rosen, M. S. (2018). Image reconstruction by domain-transform manifold learning. *Nature*, *555*(7697), 487–492. https://doi.org/10.1038/nature25988