Interpretable AI in Healthcare: Enhancing Fairness, Safety, and Trust

Samual MacDonald, Kaiah Steven, and Maciej Trzaskowski

It is becoming evident that AI has a great potential to improve the healthcare industry and the overall health and wellbeing of people. However adoption into clinical practice still faces the hurdle of a "lack of transparency" [34]. AI assisted healthcare services lack transparency when one cannot *interpret the reliability or reasoning* of the AI model involved with decision making. Without *interpretable AI* a practitioner cannot explain results to clients and is therefore unable to guarantee safety. Safety is a key element to gain trust from both clinicians and patients, especially for life-critical decision making. Thus, interpretable AI is key for the successful and safe adoption of AI in healthcare. As such, in this chapter we aim to elucidate what interpretability means and why most machine learning (i.e. AI) models fail to satisfy these definitions. We do this by propositioning a layered structure of Interpretable AI (see Fig. 1), whereby the ordered layers of uncertainty, significance, and causality have increasing amounts of "*interpretive power*" [23]. We first detail this layered structure to define interpretability and then explain how it interplays with the so-called "black box problem" of AI. Then we peel back each conceptual layer of interpretable AI in turn, while describing their respective applications and limitations. First, we explain how the foremost layer of uncertainty is key for safety and fairness by illustrating applications in supporting decision making and clinical trials. Second, we address how "significance" in AI may assist in determining the input features contributing most towards a given prediction. Finally we arrive at the central core of interpretable AI of "causality", which is key to answering "what if" type questions that are key to estimating treatment effects, treatment recommendation, and improving clinical trials.

S. MacDonald · K. Steven · M. Trzaskowski (✉)
Max Kelsen, Spring Hill, QLD, Australia
e-mail: maciej.t@maxkelsen.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

M. Raz et al. (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-981-19-1223-8_11

1 Introduction

It stands to reason that being able to Interpret AI (e.g. machine learning) in healthcare supports clinical decision making (e.g. diagnosis or treatment recommendation). The "interpretive power" of a model will dictate criteria relating to *safety*, *reliability*, and *fairness* as illustrated in Fig. 1. These criteria cannot be guaranteed with *interpretations from predictions* alone, which have limited "interpretive power". Predictions alone provide no intel about the reliability (likelihood of correctness), nor do predictions inform us about how models make their conclusions. For a more comprehensive inference we also need *interpretations from uncertainty* to flag false predictions and know what we don't know so that we can communicate explanations about *safety*. For an even stronger interpretive power we need *interpretations from significance* to identify the patient "input" variables (or factors) that contribute most strongly with a correct prediction. *Interpretations from causality* provide the most interpretive power by allowing us to estimate necessary treatments; improve fairness by controlling for sensitive variables such as ethnicity, sex, and religion; and to avoid confounding biases. All of these interpretations are used to help elucidate the choices, architectures and settings of the models employed, to in turn combat the black box problem (more on the black box problem below).

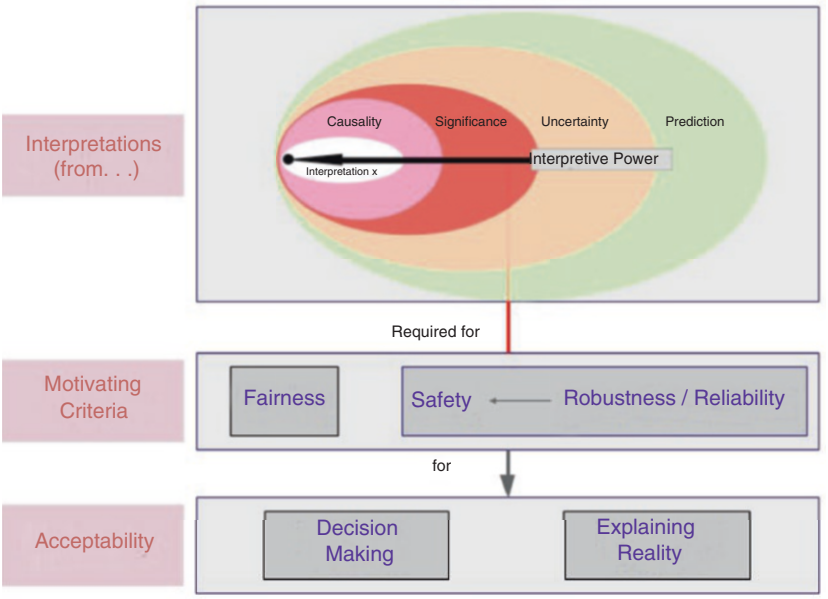

Fig. 1 The canonical structure of quantitative interpretations [23]. All models are equipped with predictive capabilities, the foundation of uncertainty, which in turn informs significance and then causality. With causality, arguably possessing the highest data-driven interpretive power, we may answer counterfactual "what if" and "why" type questions, and begin to reason further away from training examples (a new frontier for statistics and machine learning)[3]

1.1 Defining Interpretability

Interpretability is notably quite difficult to define exactly. It often depends on domain specific context to fully express and is inextricably tied to the medium that does the interpreting (e.g., clinician, patient, AI model). Throughout this chapter when we talk about interpretability, we specifically refer to what is interpretable to a *human*. If an interpretation makes sense to us, it may improve our current understanding of the world (which does not instantly equate to a universal truth). Framed another way, this can be thought about as *domain-specific* knowledge transfer. If a model provides information which can be integrated into a new domain-based understanding, then it is interpretable. For example, binary machine code is entirely interpretable between computers, but to relay information to humans it must be translated into an interpretable domain through sound and sight. The same can be said about interpretable AI, the goal is to convey valuable information to us using a medium we know and understand. This allows the models to become generators of new comprehensive information, rather than just being devices that process information. For example, if a model developed to detect rare cancers was interpretable, it might be capable of constructing a hypothesis about what factors contribute most significantly towards the cancer's progression, thus highlighting viable candidates for future interventions.

Another closely related concept worth mentioning is *explainable* AI, often used interchangeably with interpretability by many authors. Throughout this chapter we refer to explainability as an *interpretation that's tailored to a specific end user,* which reveals information about how a model reaches its output. For an explanation to be practical it must be faithful to the actual model, comprehensive to the end user, and complete with respect to the dynamics of the entire model, i.e. the explanation would provide enough information to compute the output for *any* input within sensible bounds. We note that all explanations are a form of interpretation, however not all interpretations serve as explanations. We use the terminology interpretability throughout the document as we note that some interpretable AI methods do not provide full explanations but are still useful in pursuit of the same motivating criteria.

1.2 The Black-Box Problem

The blackbox problem is an issue related to the *opaqueness* of learning processes within machine learning (ML) models and is particularly relevant to deep learning models. The problem is that, while ML often achieves superior *predictive* performance in very narrowly defined tasks, ML solutions tend to fall short of being sufficiently interpretable. As stated, this is especially problematic in high-risk domains like healthcare, where decisions cannot be treated lightly as they often affect a patient's wellbeing. Ascertaining exactly how the model achieved its performance is

still a major open question and an area of active research. The more complex questions we ask the more complex models we build, making it increasingly difficult to know how the models work**,** rendering them opaque, hence the term "black box".

A class of ML particularly susceptible to the black box problem – because it is increasingly trending towards higher model complexity – is deep learning. Deep learning has made evident that there is a strong positive relationship between state-of-the-art *predictive* performance and the number of model parameters. Take for example the full version of Open-AI's powerful natural language processing model, GPT-3, which has about 175 billion learnable parameters [6]. A price for the complex structure inherent in deep learning architectures is the consequential difficulty in making interpretations about uncertainty, significance, and causality with sufficient fidelity.

This highlights a commonly perceived trade-off between interpretability and predictive capabilities of models. Which is not to say high fidelity interpretations are *impossible*, but rather the difficulty of interpreting goes up with model complexity, a rather straightforward outcome. As an example, the humble linear model – so ubiquitous throughout the sciences – is so appealing because of its simplicity and hence interpretability. It allows for transparent interpretations about the significance of input variables (i.e. features), as well as offering tractable uncertainty measures. Conversely, neural network models often host many compositions of multivariate and nonlinear functions, which are not directly interpretable.

The black box problem, i.e. the difficulty interpreting complex models is a major hurdle impeding the common utilisation of deep learning models throughout healthcare. So how do we overcome the problem and enjoy the benefits of AI in healthcare?

Throughout the following sections, we will discuss how the black box problem can be addressed, by targeting the individual layers and components of interpretability, as well as additionally outlining the limitations of these components. While there are no universally accepted definitions of interpretable AI within the literature, we posit a structure of "orders of interpretive power" described in Fig. 1. Ordering the layers from least to most interpretable, as: uncertainty, significance, and causality.

2 Interpretations from Uncertainty – Explaining Reliability for Safety and Trust

The safety of AI in healthcare depends on being able to communicate AI reliability with good uncertainty quantifications [3] . Many things may be communicated about as there are various sources of uncertainty. Such sources of uncertainty *non-exhaustively* include (1) data acquisition error, (2) sub-population under-representation, (3) predictive error, (4) model suitability, (5) randomness inherent in the modelled phenomena, and (6) presence of hidden confounders [3]. Uncertainty

estimates may or may not represent these sources of uncertainty so we can exploit them and improve safety.

2.1 Applications of Uncertainty

Uncertainty Is Often Used to Manage the Inherent Risk in AI Systems Consider the example where genomic information is used by an AI model to predict whether a patient responds positively to an extremely expensive and high-risk intervention – treatment X. A doctor may use this "outcome predictor" model to support their decision making. Now imagine an unlucky patient is subjected to an incorrect prediction from the AI model, incorrectly forecasting how the patient will respond to treatment X. If the doctor decides to rely on this AI model, using only the predictive information (ignoring uncertainty), the patient would undergo treatment, suffer financial stress, biological side-effects, without any positive impact from the treatment. On the other hand, if the doctor chooses to inspect the AI model's uncertainty estimate, they may notice the uncertainty is unacceptably large, deeming the AI prediction unreliable, and reject the prediction. Consequently, utilising uncertainty measurements allows for *risk management,* facilitating a safer and more trustworthy decision-making process.

This "risk management" process can be scaled up to systems that make more frequent predictions, whereby predictions about some calibrated uncertainty thresholds are deemed unreliable and rejected semi-automatically. Machine learning engineers use such risk management processes often, usually calling it "uncertainty thresholding" or "risk shedding". It is important in this application that engineers ensure stakeholders are informed about what percent of data are rejected (at the class level). This is important because sometimes risk shedding will first discard under-represented sub-populations (as they have higher model uncertainties).

Uncertainty Can Improve the Safety of Clinical Trials A recent proposal [20] illustrates this in the context of phase one clinical trials, which aim to determine the maximum tolerated dose of a drug, constrained by maintaining levels below some acceptable toxicity. These trials can be adaptively proceeded, starting with low drug doses with negligible toxicity. The drug dose is then slowly increased or decreased, depending on the frequency of patients who experience toxicity levels above the acceptable safety margin. Lee et al. [20], proposed a way to improve the safety of this procedure with an uncertainty aware model. The model is uncertainty aware because it estimates a *distribution* of toxicity levels as a function of drug dose (and all past observations). By employing uncertainty, i.e. estimating a distribution of toxicity levels, the largest dose increase (constrained by confidence) can be estimated so that the toxicity levels remain within an acceptable safety margin. This is because the dose increase was constrained by confidence (a common notion of uncertainty).

Uncertainty Has Many Other Applications Lee et al. [32] Lab showed how uncertainty may support the discovery of sub-populations in clinical trials, while maximising heterogeneity between groups and homogeneity within groups. A specific type of uncertainty, "model uncertainty" correlates with the sample size of classes (or similar data points) which allows for a plethora of techniques to represent under-represented sub-populations more fairly [16]. Additionally, model uncertainties can indicate there are too little data on record to rely on the AI system, which can help inform doctors when additional laboratory testing is needed. "Distance aware" uncertainty quantification techniques [18, 28] can also flag whether the input data is not supported by the training data (i.e. too different) for the modelling predictions to be relied on (something other uncertainties struggle with). There are many more applications of uncertainty.

The above examples illustrate some convincing applications of uncertainty that improve safety in AI assisted healthcare services. But of course, this all runs on the assumption that uncertainty estimates are perfectly reliable, which of course, they are not. So next we detail the basic limitations of uncertainty, to help understand when AI is overconfident and therefore dangerous.

2.2 The Limitations of Uncertainty

The major caveat to using uncertainty to proxy the predictive reliability is that modern machine learning models struggle to provide quality uncertainty estimates. This is mostly due to modern machine learning models (e.g. deep learning) being highly parameterised (like deep learning) [12] while failing to generalise to data distinct from the training data [7]. Importantly, most machine learning algorithms are deterministic in order to make them scalable to large datasets. Deterministic algorithms often ignore the inherent randomness of nature, which is a major source of uncertainty (as discussed above). To overcome this limitation, uncertainties can be estimated by "randomising model outputs"[1] (e.g. by randomising the parameters with "Bayesian inference"; see Fig. 2). While radomising the model outputs is not necessary to approximate uncertainty, there is plenty of theoretical and empirical evidence suggesting that such randomised AI enhances the reliability and therefore safety of AI [14, 17], which is absolutely imperative in healthcare's life-critical domain.[2]

There are still major hurdles to overcome before obtaining reliable uncertainties in AI and especially deep learning. One major issue is that a model's uncertainty estimates fail to generalise beyond what the model is experienced with. A common failure mode if test data are too different from training input data (e.g. if the data

[1] We abuse terminology here as the details are esoteric and a little beside the point. Please refer to please refer to MacDonald [24] for a more accessible details about uncertainty in deep learning.

[2] See Davis et al. [14] for a practical prescriptive guide on how to estimate uncertainties in deep learning.

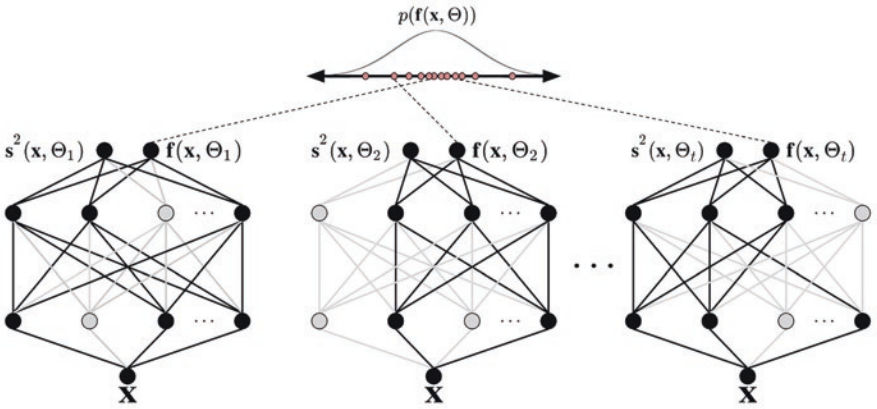

Fig. 2 Monte Carlo dropout is one way where model outputs can be randomised. Given a single input x, the neural network model makes multiple different predictions of $y = f(x, \Theta_t)$ of by switching off (grey circle) and on (black circle) neurons at random. The uncertainty is simply the variance (i.e. "spread") of all these estimates. (Figure extracted from Ref. [24])

belongs to an unseen and novel class) is that extremely small uncertainties will accommodate false predictions, which we call "Silent Catastrophic Failure". This is in large part because a central assumption is broken: test data are not independent nor identically distributed to the training data. For example, some models trained on data from one hospital can have poorer accuracy when testing on data from a different hospital, while additionally yielding misleadingly smaller estimates of uncertainty [14, 25]. This is a heavily studied pathology of AI which is often called *overconfidence*. This overconfidence problem has many contributing factors, some of which are caused by test data being too different from training data, but there is also contribution from the design choices of deterministic models. It has been proven that passing a single point estimate through the SoftMax layer of a neural net will lead to larger confidence scores than if a distribution is parsed through the SoftMax layer [17]. Overconfidence can lead to unsafe predictions and a higher rate of false predictions, which in healthcare is unacceptable. Research that aims to overcome this specific pathology suggests that in big data settings deep ensembles [11, 18] and "distance aware" models (Amersfoort et al. [39, 40]) might be the most practical and most effective way forward.

3 Interpretations from Significance – Explaining Predictions with Associative Reasoning

Almost all machine learning problems are framed with respect to some set of inputs (i.e. features) which are then transformed by the model to produce the desired output (e.g. classification). Ultimately, each of the features used will have some level

of contribution towards the model's output. However, identifying the most important features – those that contribute most towards inference – requires the next step in interpretive power: significance. Some features contribute more meaningfully to the task at hand than others, both at a global scale (across the whole population of data points) and a local scale (about individual data points). Interpreting feature importance from machine learning models is not a trivial task as the models are non-linear and a high dimensional parameter space must be translated into some measure of feature significance. There are a number of different statistics available to help tackle this (e.g. SHAP [22], LIME [29], L2X [8], and p-values [31]), each with their respective strengths and weaknesses. However, methods to obtain such statistics (except for p-values and Bayes factors, the latter of which we do not discuss herein) are reliant on interpreting feature importance by simple ordering on the magnitude of their respectively calculated numerical scales, saying nothing about how far down the list the meaningful features propagate, resulting in arbitrary cut-offs at a desired feature number.

This limitation stems primarily from the fact that the distributional properties of the saliency estimates, such as SHAP values, have not yet been mathematically proven. Consequently, preventing calculation of thresholds trusted by frequentists, such as p-values. Although there are potential avenues to overcome this limitation, there is no doubt that the lack of established significance standards makes it very difficult to reliably separate signal from noise. One simple way of approximating p-values would be by numerical ranking of the values converting them to quantiles drawn from some assumed, underlying distribution (the most frequently assumed is gaussian). However, just as it is true with the distributional transformations of data, e.g., taking the log of variables to improve the assumption of normality, it does not change the fact that the distribution of the original values could have been very different from the assumed ranks, making interpretations of significance tricky and unsafe.

Understanding the nature of the phenotype of interest can also present an alternative solution to p-values. For example, Yap et al. [42], investigated reliability of SHAP values, by benchmarking the results from a Neural Network (NN) against a known and trusted traditional bioinformatic tool, edgeR [31]. The authors used *IntegratedGradient* method from the SHAP family, which is essentially an amalgamation of Integrated Gradients [38], SHAP [22], and SmoothGrad [37] to explain learning processes of the NN. Given that their trait of interest was predicting differential expression of multiple tissue types, they estimated a point of "significance" as a function of tissue exclusivity, drawing the line at 50% of identified top genes being tissue exclusive. Although conservative, this approach aimed to balance the importance of genes uniquely characteristic to an individual tissue as well the pleiotropic candidates (one gene playing an important role in many tissues) which are often observed in complex traits. This set of genes, called by the authors, "SHAP genes'', showed remarkable replicability when compared to a set extracted from a totally independent sample (using identical methodology, overalp varied between 20% and 60% across all tissues), as well as when compared to genes identified by edgeR ($r = 0.98\%$). In addition, gene-sets from within individual tissues were

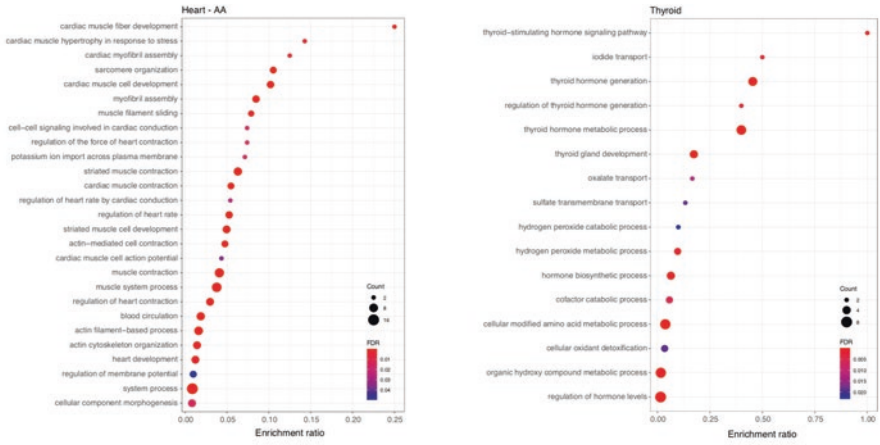

Fig. 3 Reactome pathway analysis. This enrichment of pathways for Heart – AA (atrial appendage) and Thyroid tissue. (Figure extracted from Ref. [42])

enriched in pathways consistent with expectations from domain experts in biology (e.g., Fig. 3).

Further benchmarking against edgeR, a more standard (and trusted) technique, revealed that expression profiles of the *SHAP genes* could predict tissue types more accurately than *edgeR genes* even though there were fewer of them (n = 2423 and n = 7854 respectively). Prediction accuracy in this instance was inferred from the quality of cluster formations in an independent dataset. Clusters were calculated using k-means clustering on the UMAP dimensionality reduced data, followed by the calculation of the V-Measure, which is essentially a measure of not only how well samples of the same label group together but also how homogenous each cluster is (refer to Fig. 4).

As mentioned earlier, the cut-off threshold was selected to optimise the balance between unique gene signatures within as well as across tissues. This design was purposeful as edgeR is limited to pairwise comparisons, hence designed to identify the signature of an individual tissue separately. The comparison of the effect sizes estimated for these genes showed that NNs are indeed sensitive to subtle effects of those general genes (Fig. 5).

This study illustrates that although the lack of p-values makes for challenging interpretability of AI, alternative ways of identifying "significant" features exist and can be very useful if designed and tested very carefully.

Finally, the difficulty with many learning features is they are fundamentally based on correlations between the input features and output features, so they don't *necessarily* tell us anything about causality (which will be discussed later). However, as demonstrated above, knowledge of feature contribution does enable a greater interpretation of the model's inner workings. Illuminating exactly which pieces of information *stand out*, even if at an arbitrary level, can be used not only to identify important features, but also to prevent silent failures due to ever present

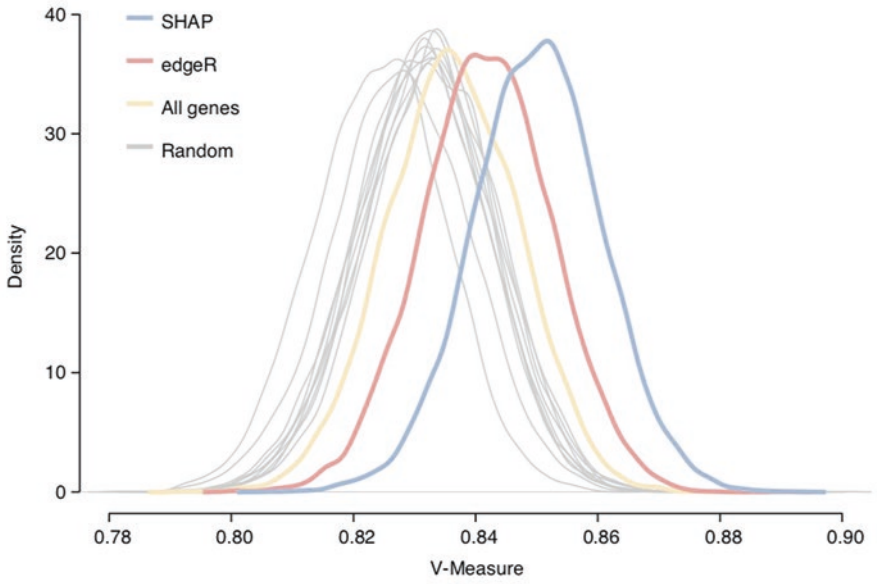

Fig. 4 V-Measure of k-means clustering analysis on UMAP. k-means clustering was performed using SHAP genes (2423 genes), 10 random sets of 2423 genes, edgeR genes (7854 genes), and all genes (18,884 genes). (Figure extracted from Ref. [42])

confounders. As an example of such danger, an AI algorithm used to diagnose lung infection from an X-ray scan was found to use the type of X-ray machine used, when making its predictions [10]. A closer look revealed that patients who were too sick to get out of bed were X-rayed using a portable type of x-ray, hence the *spurious correlation* that should have been controlled for. This model biasing could have potentially catastrophic outcomes if its predictions were trusted in real-world scenarios.

4 Interpretations from Causality – Explaining Hypothetical Outcomes with Causal Reasoning

Interpretations of causality are the most informative for explanations. As mentioned above, most ML methods only recognise associations between variables (e.g. linear correlations). Associative models disregard causality and as such are unable to disentangle confounding biases or ask hypothetical "what if" kinds of questions. Conversely, models endowed with the correct and *complete* causal structure possess the direction between causes and effects, thus enabling causal reasoning. Causal reasoning involves estimating different outcomes from corresponding hypothetical interventions which may or may not have yet been observed. This ability to estimate

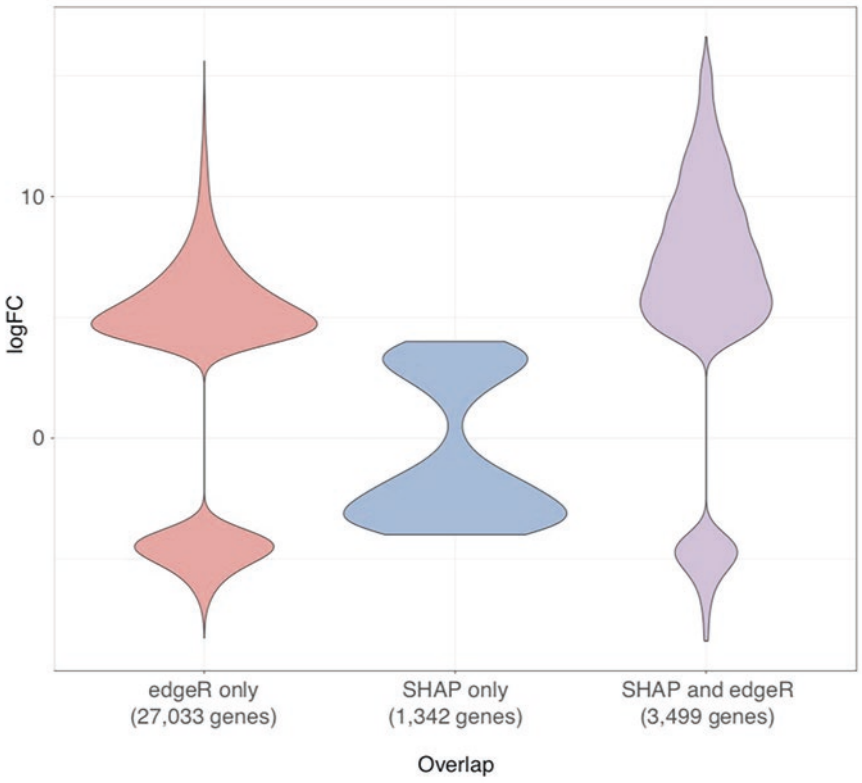

Fig. 5 Violin plots of the top genes unique and common to the edgeR and SHAP methods at the individual tissue level. (Figure extracted from Ref. [42])

outcomes, or "imagine", via correct causal reasoning would improve control in decision making processes. For example, intervening with variables (e.g. levels of a drug) that are causal to an outcome of interest (e.g. toxicity) will affect patient outcomes. On the other hand, intervening with variables (e.g. alertness caused by the drug) that are not causal to the toxicity will have no effect, even if observational records show a *spurious* correlation between alertness and toxicity. This example is just a thin slice of why causality is important. The whole picture is far more complicated, and is still a very new frontier in machine learning research.

There are two major fields of research in causal inference: (1) estimating causal relationships and (2) estimating causal effects. In this chapter we only address estimating causal effects.[3] For the rest of this section, we outline why we are motivated to understand the role of causal machine learning using examples in diagnosis, and

[3] We refer the interested reader to Peters J. et al. [27] for a comprehensive review of causal inference including methods for estimating causal relationships.

treatment recommendation, while addressing the key assumptions and limitations that must be addressed.

4.1 Diagnosis

Causality Is Crucial for Explaining Risk Factors in Diagnosis This can be highlighted by considering the example detailed by Richens et al. [30], where an AI model was tasked with diagnosing pneumonia. In this example, the level of patient care was not accounted for and therefore a hidden confounder. Asthmatic patients admitted for pneumonia are given significantly more treatment, when compared to other patients and therefore have lower mortality rates (despite the higher risk). As such, associative models (ignoring causality) would figure asthma as a protective risk factor. This erroneous explanation would be a dangerous conclusion as the ML model would lead an automated system to suggest less aggressive treatment for asthmatic patients. Alternatively, interpretations about the causal relationships between variables would have made such a conclusion much more unlikely, as the confounding attention could be controlled for, and asthma may not then present itself as a factor for improved outcome.

Causal Reasoning Can Improve Predictive Accuracy Association based models may leverage spurious relations that would be damaging when evaluating data distinct from what it was trained for (e.g. testing in a new Hospital or country). Causal models suffer less from this pitfall. For the task of diagnosis, Richens et al. [30] compared the predictive performances between an associative model, a causal-aware model, and a cohort of 44 doctors. The associative algorithm achieved similarly with the average doctor, respectively scoring accuracies of 72.5% and 71.4%, while the causal-aware model achieved a mean accuracy of 77.3%. The causal-aware model was particularly good at rare-diseases.

4.2 Decision Making – Individualised Treatment Recommendation

The Task of Recommending some Treatment *W* to a Patient Depends on the Description of the Patient *X* and Outcome *Y* The decision is based on the expected treatment effect T for that patient $T = Y^{(1)}–Y^{(0)}$, where $Y^{(1)}$ and $Y^{(0)}$ denote the potential outcome with and without treatment, respectively. Traditionally, treatment effects are only known at the population level (i.e. Averaged Treatment Effects) and typically quantified with careful and expensive clinical trials that can cost upwards of hundreds of millions of dollars [19]. Averaged Treatment Effects are useful but fail to differentiate effects between sub-populations of the trialled cohort

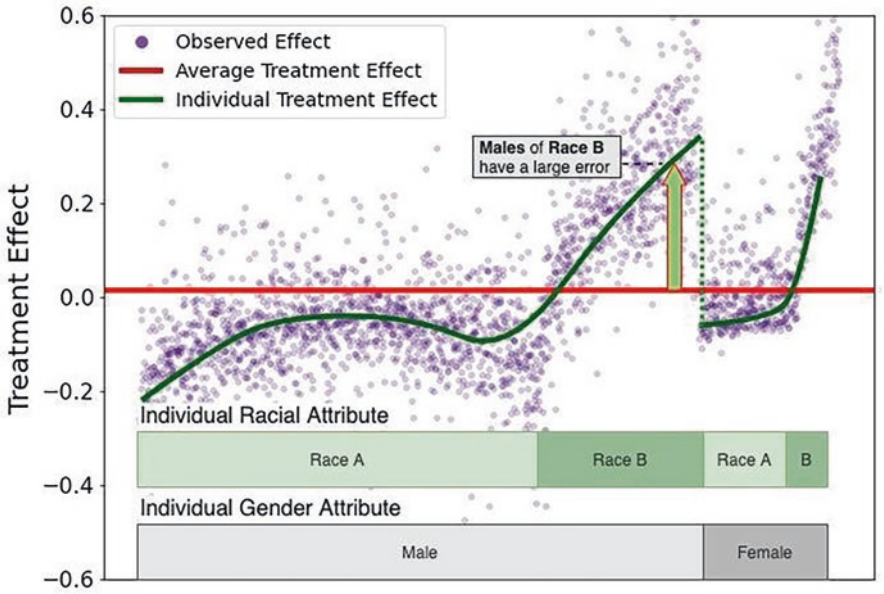

Fig. 6 Average treatment effects (ATEs) represent the average population and, if derived from a randomised clinical trial, are often trusted. ATEs do not represent the individuals, especially those who are particularly unique. This is what motivates causal machine learning techniques to model individualised treatment effects

(understand Fig. 6). On the other hand, with causal machine learning, Individual Treatment Effects can be estimated to imagine different potential outcomes $Y^{(W)}$ which depend on an individual's characteristics X and treatment option W. Thus, individualised treatment recommendations can be made by interpreting the Individual Treatment Effects [5]. There are many pros and cons for using observational data versus randomised clinical trials to estimate treatment effects, we discuss this in the next subsection. In this subsection we focus on the opportunities within learning individualised treatment effects on observational data for the task of treatment recommendation.

While It Is Possible, a Standard Supervised Model Should Not Be Used for Individualised Treatment Recommendation Treatment recommendation should be based on the expected treatment effect T. A single supervised function can be trained to learn an outcome Y based on the patient description X and treatment option W, such that $Y^{(W)} = f(X,W)$. In this scheme, the treatment effect could be approximated by toggling the treatment variable between the various states, such that the treatment effect $T = f(X, 1) - f(X, 0)$. This is not advised, as the function learns only one outcome-response curve to describe two potential outcomes. These two potential outcomes $Y^{(1)}$ and $Y^{(0)}$ may have different properties and thus need distinct ways in which they are described. To increase flexibility, other supervised

techniques aim to learn two unique functions (one for each potential outcome), but this then dramatically reduces the available learning data, which now has to be split (unevenly) between two functions. As such, causal machine learning techniques aim to learn multiple outcome-response curves using a shared function (e.g. with multi-task learning), whereby information can be shared between the two potential outcome's response curves, while respecting the flexibility required to provide unique properties to each potential outcome's underlying process. We recommend Bica et al. [4], for a good review on the topic.

Causal Machine Learning Techniques Learn Individualised Treatment Effects for a Variety of Frameworks Treatment options can be categorical, including binary treatment [1], categorical treatment [9], and combinations of treatments [4]. Each of these treatment options can be complimented with continuous dosage levels too [35]. Furthermore, side-effects can be modelled in addition to outcomes, thus enabling a compromise for a safer treatment recommendation.

4.3 Assumptions, Challenges, and Limitations in Causal Machine Learning

While there is much promise in using causal machine learning to estimate individual treatment effects and in turn allow for careful treatment recommendation systems, the modelling assumptions are complex and require careful attention in order to arrive at correct conclusions.

Observational Data for Estimating Individual Treatment Effects Is Extremely Biased (e.g Electronic Health Records) For example, patients burdened with an advanced cancer with a high tumor mutational burden may be more likely to receive immunotherapy [13]. Not controlling for this bias can lead to concluding immunotherapy treatment leads to harm. These biases additionally result in distributional differences (i.e. dataset shifts) between the treated patients and non-treated patients. Traditional supervised techniques will fail to generalise in such settings due to limitations already discussed (recall the spurious dependencies). Some ways to control for this bias can include learning balanced representations [36], or by adjustments with the propensity score (probability of being assigned treatment) [33].

Causal Inference Assumes there Are No Hidden Confounders [27] To assume there are no hidden confounders, one should consult domain experts who can be sure of its validity. This is rarely the case. As such, multiple machine learning methods have been developed to adjust for the hidden confounders. The most successful way to do this may be by modelling the hidden confounder with a "latent variable" ([21]; [41]; [43]). Zhang et al. [43] shows an example with data from an electronic health record (EHR) where non-causal treatments are recommended when latent

variable modelling is ignored, hence highlighting the significance of hidden confounders.

Causal Inference Is Only Reliable If the Potential Outcomes $Y^{(W)}$ Have Shared Support [27] In other words, the treatment effects for a patient can only be relied on if that patient has a non-zero probability of receiving any of the treatment options. This is difficult to guarantee in practice. Ways around this limitation is to either characterise the "overlap of shared support" [26]. Alternatively, we identify that the assumption of shared support may be inspected using distance-aware uncertainty estimates (e.g. from Gaussian processes), whereby large model (i.e. "epistemic") uncertainties might indicate where the assumption of shared support is not valid and thus where inference is not trustworthy.

Individual Treatment Effects Are Technically Impossible to *Perfectly* Evaluate The fundamental challenge of causal inference is that we can only observe factual outcomes. We can only observe a single outcome for each patient – the factual outcome. The counterfactual outcome that would have occurred in the imaginary world where the patient received a different treatment is unobservable. Thus, we cannot estimate the error in our estimated *individual* treatment effects. To overcome this limitation, quasi-evaluation techniques are undertaken on semi-synthetic data, whereby confounding biases are introduced by applying treatments according to unbalanced Bernoulli or Categorical distributions.

Data Quality Is the Most Limiting Challenge in Causal Machine Learning Electronic health records (EHRs) have the benefit of being highly available in large quantities and extremely heterogeneous, all of which lends itself well to data-driven techniques such as machine learning. Although algorithms trained on EHRs have less regulatory acceptance, large amounts of missing data (for many different reasons) and extreme amounts of observation biases [2]. To overcome data quality pitfalls of EHRs, efforts are being made to combine EHRs and data from randomised controlled trials (RCTs). For example, Kallus et al. [15] adjust confounding biases inherent in EHRs using data from RCTs. Other techniques can be imagined, for example we could propose to cross-check the averaged treatment effect from observation-based estimates against the average treatment effect from a corresponding RCT.

5 Conclusion

Healthcare demands transparent and interpretable AI in order for us to meet the strict criteria of safety, fairness, and reliability for acceptable decision making and justified explanations of scientific hypotheses (Fig. 1). What plagues the adoption of AI into healthcare is the black box problem, which can be explained in terms of the trade-off between model complexity and interpretability. In order to overcome the

black box problem we must be aware of different interpretations that are possible from machine learning models and their limitations. As such, we explained the canonical layers/components of interpretable AI: predictions, uncertainty, significance, and causality. We additionally explained how these different types of interpretations can support various explanations (e.g. that interpretations from uncertainty support explanations about safety and reliability). The pertinent message we wanted to convey in this chapter was that the Blackbox problem is the major barrier to reliable AI in healthcare. Once this is overcome, we expect AI to dramatically improve the productivity of the systems in healthcare while improving patient outcomes, not just in terms of their health, but also in a way that is fair to us all.

References

1. Alaa AM, van der Shaar M (2017) Bayesian inference of individualized treatment effects using multi-task Gaussian processes. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
2. Beaulieu-Jones BK, Finlayson SG, Yuan W, Altman RB, Kohane IS, Prasad V, Yu K-H (2020) Examining the use of real-world evidence in the regulatory process. Clin Pharmacol Ther 107(4):843–852. https://doi.org/10.1002/cpt.1658
3. Begoli E, Bhattacharya T, Kusnezov D (2019) The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 1(1):20–23. https://doi.org/10.1038/s42256-018-0004-1
4. Bica I, Jordon J, van der Schaar M (2020) Estimating the effects of continuous-valued interventions using generative adversarial networks. ArXiv:2002.12326 [Cs, Stat]. http://arxiv.org/abs/2002.12326
5. Bica I, Alaa AM, Lambert C, van der Schaar M (2021) From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. Clin Pharmacol Ther 109(1):87–100. https://doi.org/10.1002/cpt.1907
6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C et al (2020) Language models are few-shot learners. ArXiv:2005.14165 [Cs]. http://arxiv.org/abs/2005.14165
7. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon JV, Lakshminarayanan B, Snoek J (2019) Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.
8. Chen J, Song L, Wainwright MJ, Jordan MI (2018) Learning to explain: an information-theoretic perspective on model interpretation. arXiv. https://arxiv.org/abs/1802.07814v2
9. Chen P, Dong W, Lu X, Kaymak U, He K, Huang Z (2019) Deep representation learning for individualized treatment effect estimation using electronic health records. J Biomed Inform 100:103303. https://doi.org/10.1016/j.jbi.2019.103303
10. Couzin-Frankel J (2019) Medicine contends with how to use artificial intelligence. Science 364(6446):1119–1120. https://doi.org/10.1126/science.2019.6446.364_1119
11. Fort S, Hu H, Lakshminarayanan B (2019) Deep ensembles: a loss landscape perspective. arXiv. https://arxiv.org/abs/1912.02757v2
12. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. arXiv. https://arxiv.org/abs/1706.04599v2
13. Healthdirect H (2021) Cancer immunotherapy [Text/html]. Healthdirect Australia, September 15. https://www.healthdirect.gov.au/cancer-immunotherapy

14. Davis J, MacDonald S, Zhu J, Oldfather J, Trzaskowski M (2020) Quantifying uncertainty in deep learning systems. AWS Prescriptive Guidance. https://docs.aws.amazon.com/prescriptive-guidance/latest/ml-quantifying-uncertainty/welcome.html
15. Kallus N, Puli AM, Shalit U (2018) Removing hidden confounding by experimental grounding. Adv Neural Inf Proces Syst:31. https://papers.nips.cc/paper/2018/hash/566f0ea4f6c2e947f36795c8f58ba901-Abstract.html
16. Khan S, Hayat M, Zamir SW, Shen J, Shao L (2019) Striking the right balance with uncertainty. In: Proceedings – 2019 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2019, pp 103–112. https://doi.org/10.1109/CVPR.2019.00019
17. Kristiadi A, Hein M, Hennig P (2020) Being Bayesian, even just a bit, fixes overconfidence in ReLU networks. https://arxiv.org/abs/2002.10118v2
18. Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30. Curran Associates, Inc, pp 6402–6413. http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf
19. Ledesma P (2020) How much does a clinical trial cost? Sofpromed, January 2. https://www.sofpromed.com/how-much-does-a-clinical-trial-cost
20. Lee H-S, Shen C, Zame W, Lee J-W, van der Schaar M (2021) SDF-Bayes: cautious optimism in safe dose-finding clinical trials with drug combinations and heterogeneous patient groups. ArXiv:2101.10998 [Cs, Stat]. http://arxiv.org/abs/2101.10998
21. Louizos C, Shalit U, Mooij J, Sontag D, Zemel R, Welling M (2017) Causal effect inference with deep latent-variable models. In: Proceedings of the 31st international conference on neural information processing systems, pp 6449–6459
22. Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. ArXiv:1705.07874 [Cs, Stat]. http://arxiv.org/abs/1705.07874
23. MacDonald S (2019) Interpretations in Bayesian deep learning. University of Queensland. Master of Data Science Capstone Thesis Project
24. MacDonald S (2020) Interpretations of learning. Medium, March 3. https://towardsdatascience.com/interpretations-in-learning-part-1-4342c5741a71
25. Obermeyer Z, Emanuel EJ (2016) Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 375(13):1216–1219. https://doi.org/10.1056/NEJMp1606181
26. Oberst M, Johansson FD, Wei D, Gao T, Brat G, Sontag D, Varshney KR (2020) Characterization of overlap in observational studies. ArXiv:1907.04138 [Cs, Stat]. http://arxiv.org/abs/1907.04138
27. Peters J, Janzing D, Scholkopf B (2017) Elements of causal inference: foundations and learning algorithms. MIT Press
28. Rasmussen CE (2004) Gaussian processes in machine learning. In: Bousquet O, von Luxburg U, Rätsch G (eds) Advanced lectures on machine learning: ML summer schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4—16, 2003, Revised lectures. Springer, pp 63–71. https://doi.org/10.1007/978-3-540-28650-9_4
29. Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144. https://doi.org/10.1145/2939672.2939778
30. Richens JG, Lee CM, Johri S (2020) Improving the accuracy of medical diagnosis with causal machine learning. Nat Commun 11(1):3923. https://doi.org/10.1038/s41467-020-17419-7
31. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140. https://doi.org/10.1093/bioinformatics/btp616
32. Lee H, Zhang Y, Zame WR, Shen C, Lee J, van der Schaar M (2020) Robust Recursive Partitioning for Heterogeneous Treatment Effects with Uncertainty Quantification. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

33. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55. https://doi.org/10.1093/biomet/70.1.41
34. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x
35. Schwab P, Linhardt L, Bauer S, Buhmann JM, Karlen W (2020) Learning counterfactual representations for estimating individual dose-response curves. Proc AAAI Conf Artif Intell 34(04):5612–5619. https://doi.org/10.1609/aaai.v34i04.6014
36. Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithms. In: Proceedings of the 34th international conference on machine learning, pp 3076–3085. https://proceedings.mlr.press/v70/shalit17a.html
37. Smilkov D, Thorat N, Kim B, Viégas FB, Wattenberg M (2017) SmoothGrad: removing noise by adding noise. CoRR:abs/1706.03825. http://arxiv.org/abs/1706.03825
38. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Proceedings of the 34th international conference on machine learning – volume 70, pp 3319–3328
39. van Amersfoort J, Smith W, Teh YW, Gal Y (2020) Uncertainty estimation using a single deep deterministic neural network. Proceedings of the 37 th international conference on machine learning, Vienna, Austria, PMLR 119, 2020.
40. van Amersfoort J, Smith L, Jesson A, Key O, Gal Y (2022) On feature collapse and deep kernel learning for single forward pass uncertainty. https://arxiv.org/abs/2102.11409
41. Wang Y, Blei DM (2019) The blessings of multiple causes. J Am Stat Assoc 114(528):1574–1596. https://doi.org/10.1080/01621459.2019.1686987
42. Yap M, Johnston RL, Foley H, MacDonald S, Kondrashova O, Tran KA, Nones K, Koufariotis LT, Bean C, Pearson JV, Trzaskowski M, Waddell N (2021) Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. Sci Rep 11(1):2641. https://doi.org/10.1038/s41598-021-81773-9
43. Zhang L, Wang Y, Ostropolets A, Mulgrave JJ, Blei DM, Hripcsak G (2019) The medical Deconfounder: assessing treatment effects with electronic health records. In: Proceedings of the 4th machine learning for healthcare conference, pp 490–512. https://proceedings.mlr.press/v106/zhang19a.html