phg foundation
making science work for health

Black box medicine and transparency

Interpretability by design framework

A PHG Foundation report for the Wellcome Trust

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish, Hannah Murfet, Colin Mitchell, and Alison Hall

Acknowledgements

The *Black Box Medicine and Transparency* project was funded by the Wellcome Trust as a part of their 2018 Seed Awards in Humanities and Social Sciences [Grant Number: 213623/Z/18/Z]. We thank the Wellcome Trust for their support.

The series of reports is informed and underpinned by a series of roundtables and interviews. These roundtables and interviews are detailed in the Report of Roundtables and Interviews. Further, highlights from both are seeded throughout all reports, being found in 'A Salient Feature' boxes.

Disclaimer

The following report is intended to provide general information and understanding of the law. The report should not be considered legal advice, nor used as a substitute for seeking qualified legal advice.

Hannah Murfet (PHG Foundation Fellow) contributed to this report by way of in-kind support from Microsoft Research Ltd. Any opinions expressed are the author's own, and may not represent the view of Microsoft Research.

URLs in this report were correct as of February 2020

This report is available from www.phgfoundation.org

Published by PHG Foundation 2 Worts Causeway, Cambridge, CB1 8RN, UK

+44 (0)1223 761900

February 2020

Correspondence to: intelligence@phgfoundation.org

How to reference this report:

Ordish J, Murfet H, Mitchell C, Hall A. *Black Box Medicine and Transparency: Interpretability by Design Framework*. PHG Foundation. 2020.

Contents

Navigating interpretability

Organisations developing machine learning applications for healthcare must navigate a complex series of decisions when designing their system. One important design consideration is the human interpretability ('interpretability') of the machine learning model. That is, whether (or rather to what extent) the model is understandable to humans.[1] To illustrate, consider the example in the box below.

Example: Modelling prognostic trajectories of cognitive decline due to Alzheimer's disease[2]

Consider we want to produce a risk tool to predict cognitive decline due to Alzheimer's disease.

Suppose we have three machine learning models that have the task of classifying patients into three groups. Patients with:

A. normal cognitive function; or
B. stable mild cognitive impairment (sMCI); or
C. progressive mild cognitive impairment (pMCI) (Alzheimer's).

These models are trained using a variety of features that characterise cognitive function and impairment, including changes to specific parts of the brain derived from brain imaging (PLS-derived grey matter, β amyloid), and the presence of genetic markers (e.g. APOE 4 status).

The three models have the following (simplified) characteristics:

	Predictive accuracy (%)	**Out-of-the-box human interpretability score**
Model I	High	Low
Model II	Low	High
Model III	Medium	High

Which model should the developer favour? When (if at all) should the developer trade off predictive accuracy for gains in interpretability? How central should interpretability be for any given machine learning model?

Our hypothesis is that decisions such as these are multi-faceted, multi-disciplinary, and central to the design of the overall system. The process described in this document assists those developing machine learning for healthcare to think through interpretability with regard to their machine learning system.

This document is a distillation of findings from a wider project on interpretability of machine learning for healthcare *Black Box Medicine and Transparency*. The longer report has detailed findings on:

- The context of machine learning for healthcare and research
- Interpretable machine learning
- The ethics of transparency and interpretability
- The regulation of transparency and interpretability

We encourage readers seeking a more thorough review of these topics to read the full set of *Black Box Medicine and Transparency* reports.

This document introduces the 'Interpretability by Design Framework,' a practical tool to assist those developing machine learning for healthcare and research to think through interpretability with respect to their system.

Principles for interpretability

The Interpretability by Design Framework ("ID Framework") assists developers to think through interpretability of their machine learning models for healthcare.

The following set of principles underpin and inform the ID Framework:

1. Interpretability by design

If interpretability of machine learning systems is an afterthought, interpretability of the final product will be hampered and limited. Interpretability is a design choice, a choice that should be considered at the outset of and continuously throughout the design process - what we call *interpretability by design*.

At its most radical, interpretability may require rethinking of the computational goal of the system. A system that is designed, from the outset, to be both accurate and interpretable will be a very different system from one which methodically replaces or supplements each step of uninterpretable computation after the fact.[i] Indeed, to methodically replace or supplement uninterpretable computation after the fact is likely to be both expensive and time consuming.

2. Accuracy v interpretability

Interpretability does not necessarily come at the cost of accuracy. There is little evidence to support an inverse relationship between the interpretability of the model and the accuracy of the model.[3] However, the most interpretable machine learning model may not be the most accurate and vice versa.

3. Usability

Accuracy should not be the only metric when evaluating machine learning systems in healthcare. If the system is to interact with a human, the system must be usable. Usability is defined as a 'characteristic of the user interface that facilitates use and thereby establishes effectiveness, efficiency and user satisfaction in the intended use environment.'[4] Notably, usability may also increase or decrease the safety of a system.[5] A hyper-accurate yet unusable machine learning system represents a failure of design, as does a hyper-usable yet highly inaccurate system.

4. Interpretability as a design choice

Interpretability is a design choice not a design axiom.[6] If interpretability is important for usability or necessary for compliance with ethical and legal requirements, there is a presumption that the model should be, to some degree, interpretable.

[i] Note the parallels to the privacy by design literature: Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*. 2013; 9(3-4): 3.

In some circumstances, we may be comfortable with a degree of opacity.[7] We might have sufficient reassurance that the model produces accurate outputs or simply find interpretability superfluous for that particular intended use. In short, relative opacity can be, and is sometimes, a defensible design choice.

5. Interpretability is not binary

The question of interpretability is less a binary question: whether to make one's model interpretable or not is a question of what work interpretability might do, what problems it solves, and how interpretability might enhance the system as a whole.

The kind of interpretability method used may vary according to why interpretability is sought and the intended audience. For instance, interpretability for a developer to debug a model may look very different to interpretability in the form of an explanation to an end user. We explore this further in Step 2 below.

6. Demanding interpretability

The centrality of interpretability to any given machine learning system depends on a number of key considerations - what work interpretability performs. A low risk but highly opaque machine learning system may require interpretability but merely as a means to facilitate effectiveness, efficiency, and user satisfaction of the system. In contrast, a highly risky and highly opaque system may constitute a safety concern - interpretability here contributing to the safety of the system via usability.

Interpretability of any given model is contingent upon a number of facts. For instance, interpretability may be unnecessary to test safety and effectiveness if we have sufficient testing and auditing mechanisms in place. Nevertheless, interpretability may still be important for usability of the machine learning system.

7. Interpretability as a matter of fit

Interpretability of any given machine learning model is not a design choice to be taken in isolation; it is one design decision to be taken in concert with a host of other critical decisions. A prescription for interpretability depends on the diagnosis - what problem does interpretability solve? Interpretability sits as one method within a suite of other methods that might engender trust in a model. For instance, Ng notes that the following techniques might also be of assistance:[8]

I. *Testing and audits* - the model can be subjected to testing (perhaps in the form of general safety and performance requirements in medical device law) or audit, providing some assurances regarding its performance and safety
II. *Boundary conditions* - boundary conditions specify the range of input variables allowed, ensuring that test conditions are specified and reproducible. At minimum, impossible and absurd inputs should be recognised and reported
III. *Out of sample errors, outliers, out of distribution signalling* - out of sample errors, outliers, out of distribution instances can be recognised and reported, adding to confidence that the system is functioning as intended[9]
IV. *Gradual rollout* - rollout can be restricted to certain populations or its functionality restricted until further assurance is received

V. *Monitors and alarms* - post-release, when something goes wrong, alerts can allow us to remedy the issue and learn from mistakes

Reasons for interpretability

When considering what machine learning technique to use, developers must balance a number of elements. Apart from any accuracy versus interpretability trade off, developers must also consider ethical and regulatory reasons to render their machine learning model interpretable. Given this, there are three general classes of reasons to render a machine learning model human interpretable:

I. *Practical reasons* - related to the viability (including commercial) of the product - will healthcare professionals or patients/consumers use and rely on the model's outputs if it is uninterpretable?
II. *Ethical reasons* - there may be an ethical imperative, perhaps even a duty, to render some machine learning models for health interpretable
III. *Regulatory reasons* - regulation may require developers to render their model somewhat transparent, interpretable, or explainable

Navigating and balancing these practical, ethical, and regulatory reasons for interpretability is a vexed task. The ID Framework seeks to simplify, distil, and assist with this task.

Use case

The ID Framework assists developers and product development or research teams to think through interpretability of machine learning for healthcare, distilling many of the practical, ethical, and regulatory elements of when and why interpretability might prove important throughout the design process.

When

We recommend that developers use the ID Framework at the following points in time:

I. At the outset of the design process to consider the proposed design of their machine learning system, adjusting accordingly
II. Revisit the ID Framework at key stages in the design process to consider how the evolution of the system and new information has changed the system's score
III. Post-market in response to possible changes in response to user feedback, adverse events, and to consider proposed changes to the system

Who

Many of the judgments considered as a part of the ID Framework and the decisions made as a consequence will require input from a multidisciplinary team that might include:

- Machine learning engineers and researchers
- Software development engineers
- Product managers
- Human factors experts
- Clinical expertise
- Regulatory and compliance managers as well as quality assurance professionals
- End user feedback e.g. from clinicians or patients
- Marketing managers
- Privacy professionals

Key questions

Developers should consider two key questions with respect to interpretability and the intended use of their machine learning system while using the ID Framework:[10]

I. *Verification* - broadly captured by the question: are we developing the system right?[11]
II. *Validation* - broadly captured by the question: are we developing the right system?[12]

Designing interpretability framework

The ID Framework has seven main steps to assist developers in thinking through the interpretability of their system. Notably, the Framework does not seek to list desirable or undesirable traits of machine learning but consider attributes of machine learning systems and their relation to interpretability.

Step 1: Assessment information

Step 2: Targeted interpretability, tailored explanations

Step 3: Score the interpretability of the proposed or existing model

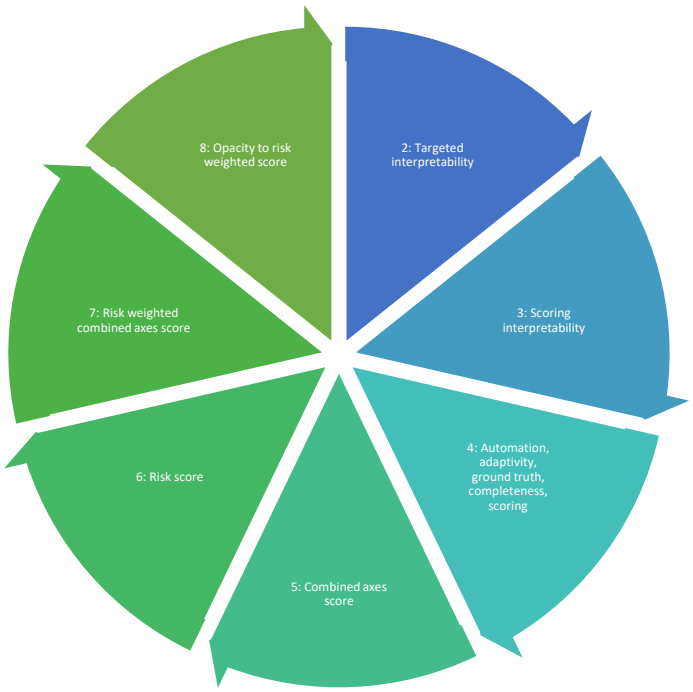
Step 4: Score the model according to the axes of automation, adaptivity, risk, (lack of) ground truth, and (in)completeness

Step 5: Generate a combined axes score

Step 6: Score the riskiness of the system and its intended use

Step 7: Multiply the combined axes score by the risk score to obtain a risk weighted combined axes score

Step 8: Combine the interpretability score with the risk weighted combined axes score to consider the calculated ratio



We provide a step-by-step guide to the ID Framework below.

How to use the ID framework

The ID framework should be used in the following way:

- You should score your system, not necessarily relative to other similar systems, but according to an objective scale. For instance, an acute triaging tool may be less risky in comparison to other acute triaging tools but this type of device may be inherently risky. This aside, you might use other similar systems as reference points to establish a score.
- You should score your system as a whole with respect to the multiple functions, tasks, or decisions it assists or completes.

We use these terms in the following way:

- *Feature* = 'Features are the input variables to a machine learning model. For example, when developing a model predicting stroke risk, a feature would be a patient's height or weight. Features can be processed before they are entered into a model, such as combining height and weight into a body mass index. For an image, a feature may be some component of the image, such as an eye or a nose, when developing a facial recognition machine learning system.'[13]
- *Machine learning algorithm* ("algorithm") = dictates how the model is trained, how the features are structured with respect to the task. For instance, convolutional neural networks in the context of image classification dictate how the model learns edges and distinguishes between classes of images.[14]
- *Machine learning model* ("model") = is 'produced as the output of a machine learning algorithm applied to training data.'[15] Sometimes also called 'the trained model,' the model here has been trained according to the machine learning algorithm on a training set of data.
- *Machine learning system* ("system") = the device which encompasses the *machine learning model*. The wider system might include the user interface, supporting architecture, visualisation of the model, as well as any physical device in which the software is embedded.

Use of symbols

- ★ The star symbol indicates specific actions for the reader to undertake
- ➢ The arrow symbols provides further clarification

★ **The following sections may be printed and used as a template to aid discussion and application of the ID Framework**

Step 1: Assessment information

This information should provide relevant context to assessing your system.

★ Fill the details of your system and assessment team in the box

Name of system: _____

Date of assessment: _____/_____/_____

System version/release candidate number: _____

Assessment participants:

Name	**Role**
e.g. Jane Smith	e.g. Senior Developer

Changes since last assessment:

Change	**Planned impact upon interpretability**
e.g. Post hoc interpretability method implemented in the form of semantic map	e.g. Clinicians may now sense check the findings of the model following the semantic map

Comments on assessment information:

Step 2: Targeted interpretability, tailored explanation

As a preliminary exercise, it is important to have concrete ideas about:

I. *Why* might you render the model interpretable?

For instance, is the primary benefit of interpretability to debug your model? Is interpretability necessary for users to contextualise its outputs? Is interpretability necessary to provide sufficient clinical evidence under the Medical Devices Regulation (MDR) or *In Vitro* Diagnostic Medical Devices Regulation (IVDR)?[16] Is interpretability necessary to understand the generalisability of the model or a helpful attribute to test for bias?

II. *Who* is interpretability or explanation for?

Consider the various purposes of interpretability. Who is the explanation for? How should the explanation be communicated to this audience?

There are at least six key audiences for interpretability or explanation in the context of healthcare or research (see Table 1 below).

Table 1: key audiences for interpretability

Audience	**Primary purpose**	**Example**
Developers themselves	To debug, understand the behaviour of, and iterate on their model. To verify, validate, and properly label the system	Semantic maps informing predictions based on automatic analysis of radiological images might assist in picking out confounding factors[17]
Regulatory bodies	To evidence the safety and effectiveness of the device. For instance, according to the MDR/IVDR and associated harmonised standards	Intrinsically interpretable machine learning models or methods that transform models into decision rules (for example, RuleFit) may make it easier to link a model's reliance on features with supporting scientific literature[18]
Commissioning bodies	To evidence the system as a cost effective tool for use in a health system. For instance, the National Institute for Health and Care Excellence (NICE) Evidence standards	Global interpretability in the context of discharge management tools might demonstrate both conformance to policy but also demonstrate the return on investment for such a system. For example, if/once it is

	framework for digital health standards[19]	known why the tools recommend discharge, we know the optimal patient pathway to follow
Healthcare professionals	To contextualise and interpret the output to make a clinically relevant action To contextualise and interpret the output for their patient To evidence the safety and effectiveness of the device	Global interpretability to understand what features the model generally finds significant to link to the healthcare professional's clinical judgment Local interpretability to understand what the model found significant for a particular patient
Health consumer / User	To contextualise and interpret in order to take an action related to their health or care To contextualise and interpret for themselves the outputs of the model To consider the system reliable or safe for their own use	Global interpretability to understand what features the model generally finds significant to link to the user's own understanding Local interpretability to understand what the model found significant for that particular user
Public	To assist in the public justification for the deployment of a system	Global interpretability to understand what features the model generally finds significant to consider the acceptability of reliance on these features Local interpretability to allow a human in the loop to be an effective checker
Knowledge discovery	To assist in scientific discovery or assist in establishing causation	For example, partial dependence plots show the marginal effect a feature has on the predictive outcome.[20] Accordingly,

	To ensure conclusions of studies including machine learning are reproducible and appropriately benchmarked	they may be useful for contextualising and interpreting a model

Apart from primary audiences listed in the above table, there are a number of secondary audiences that may have distinct purposes for interpretability. For instance, interpretability required for a product liability or clinical negligence lawsuit might be more forensic in its requirements than other audiences mentioned above. Further, investors and venture capitalists may also find interpretability of assistance when considering in which system to invest.

III. *When* is interpretability or explanation required?

Consider the various purposes and audiences for interpretability. When should the explanation occur? Should the model be interpretable before use, does the explanation arrive just in time, after use to contextualise outputs?

- ★ Outline in the below table the key purposes for interpretability, the primary audiences for that purpose, and when the interpretability is required for each

Key purpose	**Key audience(s)**	**When in time?**
e.g. To debug the model	e.g. A. Developer B. Documented as a part of medical device compliance	e.g. A. Throughout development and post-market surveillance B. Primarily before submission

Step 3: Scoring interpretability

Score your model according to its opacity. Opacity is defined in the box below.

OPACITY

How uninterpretable is the machine learning model?

Opacity is defined as the lack of interpretability of the machine learning model.

Interpretability is defined as 'the ability to explain or present in understandable terms to a human.'[21]

Note: machine learning systems may be black boxes for many reasons. For instance, Burrell, distinguishes between three forms of opacity:[22]

I. Opacity as intentional corporate or state secrecy
II. Opacity as technical illiteracy
III. Opacity that arises from the characteristics of machine learning algorithms and the scale to apply them usefully

The opacity we refer to is closest to III. - the interpretability of the model, not the intentional restriction of information (I) or technical illiteracy of the intended audience (II).

Interpretability refers to both *global interpretability* and *local interpretability* of the model.

Global interpretability is defined as understanding of 'the whole logic of the model', the ability to follow the 'entire reasoning leading to all the different possible outcomes.'[23]

Local interpretability is defined as understanding of the reasons for a specific decision - the 'single prediction/decision.'[24]

When scoring interpretability, note:

- Interpretability of the model should be assessed with respect to the key purposes and audiences you outlined above
- Evidence of interpretability might include functionality-grounded evaluation, human-grounded evaluation, and application grounded evaluation (see Interpretable Machine Learning report)[25]
- If you at the outset of product development, consider the out-of-the-box interpretability of the proposed model
- If you have used this framework before and have implemented visualizations, post-hoc explainers, or used other methods to assist in rendering your model interpretable, score the model's interpretability with these in mind. This framework might provide an assessment of the added benefit of these methods.

Scale

1 - Readily human interpretable

2 - Some elements are human interpretable

3 - Few elements are human interpretable

4 - Not human interpretable

★ Score your system (overall) by selecting the most appropriate score in the box below

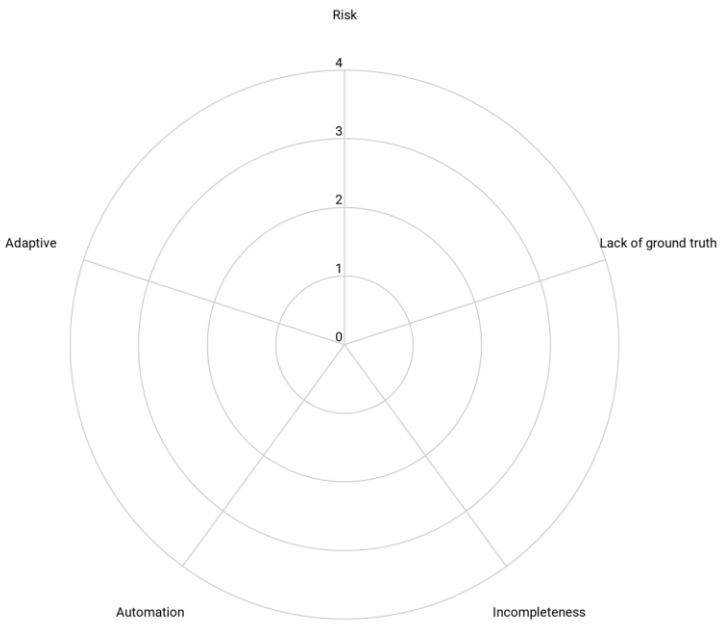
1	**2**	**3**	**4**
Interpretable			**Uninterpretable**

➢ Even if your system is relatively interpretable, developers should continue through to the tool to understand the importance of interpretability with respect to your system

Step 4: Scoring axes

The radar diagram for interpretability by design has 5 axes: automation, adaptivity, risk, ground truth (lack of), and (in)completeness (see Figure 1 below). Step 3 asks you to score your system according to all axes *apart from* risk. You will be able to plot your system on the radar diagram in Part 4 of the interpretation of scores below.

Figure 1



Each axis is elaborated on below.

Axes explained

This section further explains each axis, providing a description of the axis, indicating how the axis relates to interpretability, and what a score from 1-4 might look like.

AUTOMATION

How automated is the decision or task that the machine learning system makes or assists with?

Automation is defined as the extent to which the machine learning system contributes to the decision or task based on its intended use.

- Think about the machine learning system and how it fits into the workflow - how much and what significance does the human have in the decision/task loop?
- For example: does the system automatically generate its outputs? Does the system automatically generate its outputs but requires clinical interpretation to action? Does the system automatically make and deliver a diagnosis to a patient? Does the machine learning system automatically deliver some clinically relevant action?

Scale

1 - the decision/task is primarily human determined and actioned, the system being one of many considerations

2 - the decision/task is human determined and actioned but significantly informed by the system

3 - the decision/task is determined and/or actioned by the system but a human is in the loop

4 - the decision or task is a closed, automated loop with no human intervention by default

★ Score your system by selecting the most appropriate score in the box below

1	2	3	4
Primarily human determined/action			**Primarily machine determined/actioned**

ADAPTIVITY

How adaptive is the machine learning algorithm?

Adaptivity is defined as the frequency with which a machine learning model retrains.

We distinguish between:[26]

I. A model's inputs being updated in response to new data - while the data input changes, the model remains the same
II. A model that retrains. In this case, new data become new training data

Models that retrain (II.) are considered 'adaptive' in this context.

There are different shades of adaptivity, these shades include:

"Locked" or "static", meaning retraining either does not occur or is actioned through

planned change management principles, following a formal release process.

"Batch learning", meaning retraining of the model occurs in batches, that is, 'every so often' in accordance with new data.[27]

"Incremental learning", meaning retaining of the data occurs whenever new data is encountered.[28] If the system imports streaming data, retraining occurs continuously.

The adaptivity of a model is an engineering choice. In principle, most machine learning algorithms are capable of retraining models in response to new data. Consequently, the same machine learning algorithm may be adaptive or static depending on what is decided.

Scale

1 - static, locked

2 - infrequent batch learning

3 - frequent batch learning

4 - incremental learning using streaming data

★ Score your system by selecting the most appropriate score in the box below

1	**2**	**3**	**4**
Static model			**Incremental learning using streaming data**

INCOMPLETENESS

Are all relevant features included?

Completeness is defined as the inclusion and quantification of all features relevant to the decision or task the system's intended use references.

Incompleteness is distinct from uncertainty.[29] Uncertainty can be included in a machine learning model as a formalised, quantified value. For instance, via confidence intervals or Bayesian priors. However, incompleteness regards a relevant feature that has not been formalised, quantified, or otherwise included in the machine learning model.

"Quantification" here might include measurable observations such as blood pressure but also observable indicators such as the presence of a rash. In short, the feature is included and quantified so long as it is defined as a relevant object with respect to the model.

The assessment of whether 'all features relevant to the decision or task' will heavily depend upon the ground truth and may relate to the predictive accuracy of the model. For instance, it is only possible to say whether you have a relatively complete model if you understand how each feature contributes to the task or decision at hand i.e. you have ground truth.

A complete model may be an inaccurate model. The inclusion of all relevant features may cause overfitting. A model that has been regularised (where some features are set to zero or excluded, for example by lasso regression) may still be considered a complete model.

- Caution: the reader should note that the ID Framework does not necessarily advocate for complete models over incomplete models. The primary purpose of the tool is to consider how interpretability fits with one's model, what work interpretability might do for each model, not the desirability of any of the characteristics outlined: automation, adaptivity, completeness, and so on.

Scale

1 - confident that the model has a comprehensive feature set included and quantified

2 - many relevant features included and partially quantified

3 - some relevant features included and loosely quantified

4 - few relevant features included with weak quantification

★ Score your system by selecting the most appropriate score in the box below

1	**2**	**3**	**4**
Comprehensive feature set			**Few features**

GROUND TRUTH (LACK OF)

Is there robust identification and validation of the feature set?

Ground truth is defined as evidence that underpins the identification and validation of the feature set.

Ground truth is concerned here with a) the relevance of the feature in relation to the decision or task and b) an understanding of how each feature contributes to the decision/task. In short, strong ground truth underpins both the selection and relative prominence of features.

Notably, there is a strong relationship between ground truth and other axes, namely: completeness and opacity.

- With respect to completeness, ground truth underpins and evidences the inclusion and weighting of features in a model. Consequently, it also underpins judgments of whether the model is relatively complete or incomplete.
- With respect to opacity, ground truth often walks in lockstep with transparency. That is, if the model is opaque, the extent to which features are relied upon is hidden, ground truth cannot be applied to evidence the weighting between features. The most we can do with an opaque model is justify the mere inclusion of features according to ground truth but not their weighting.

Ground truth can take many forms but often includes the following questions familiar to those in medical device regulation:

- Is there robust scientific evidence to underpin the inclusion and, where possible, relative weighting of these features?
- Is there robust clinical evidence to underpin the inclusion and, where possible, the relative weighting of these features?

Clinical evidence in the context of a medical device might include: critical evaluation of the relevant scientific literature, evaluations of the results of all available clinical investigations, and consideration of currently available alternative treatment options.[30]

For example, consider a model to predict lung cancer risk. The model might include a feature, perhaps whether a patient smokes or not, on the basis of the scientific literature. Clinical evidence in the context of machine learning would ultimately include the predictive accuracy of that feature and how much it contributes to the overall score. In this case, it would be evidence demonstrating that the feature of smoking contributed to the overall prediction of developing lung cancer.

Scale

1- Well-established evidence for the inclusion and weighting of features

2- Some evidence for the inclusion and little evidence for the weighting of features

3- Poor evidence for the inclusion and no evidence underpinning the weighting of features

4- Evidence determines that the model relies on confounding or poorly predictive features, therefore contrary to ground truth

★ Score your system by selecting the most appropriate score in the box below

1	2	3	4
Robust ground truth			**Contrary to ground truth**

Step 5: Combined axes score

★ Calculate your **combined axes score**. The combined axes score can be calculated by the following process:

$$\textbf{\textit{Automation score}} + \textbf{\textit{Adaptivity score}} + \textbf{\textit{Lack of ground truth score}} + \textbf{\textit{Incompleteness score}} = \textbf{\textit{combined axes score}}$$

★ Record your combined axes score below

Combined axes score = _____

Step 6: Calculate risk score

Step 6 asks you to assign a score to the riskiness of the system, taking into account its intended use.

RISK

How risky is the machine learning application?

Risk is defined as the contribution made by the machine learning system to the worst-case harm arising from a hazardous situation.

"Hazardous situation" is defined as a 'circumstance in which people, property or the environment are exposed to one or more hazard.[31]

"Hazard" is defined as a 'potential source of harm.'[32]

"Harm" is defined as 'physical injury, damage, or both to the health of people or damage to property or the environment.'[33]

"Serious injury" is defined as 'injury or illness that:

a) Is life threatening,
b) Results in permanent impairment of a body function or permanent damage to a body structure, or
c) Necessitates medical or surgical intervention to prevent permanent impairment of a body function or permanent damage to a body structure.'[34]

➢ Note that evaluation of risk should occur at the outset of design, throughout the design process, and through the lifecycle of the system. Given this, you should consider *residual risk* at the latter stages of development.[35]

"Residual risk" is defined as risk remaining after risk control measures have been implemented.'[36] For instance, where a model requires input from a user, setting boundary conditions mitigates against wildly inaccurate inputs resulting in inaccurate outputs. For example, excluding or alerting a user where they input a weight of a human being over 500 kilograms.

➢ For more information see ISO 14971 and BS EN 62304

Scale

1 - the machine learning system cannot foreseeably contribute to a hazardous situation resulting in injury

2 - the machine learning system may contribute to a hazardous situation resulting in a non-serious injury

3 - the machine learning system may contribute to a hazardous situation resulting in a

serious injury

4 - the machine learning system may contribute to a hazardous situation resulting in a serious injury and threat to public health

★ Score your system by selecting the most appropriate score in the box below

1	2	3	4
No hazardous situation			**Risk to public health**

Step 7: Weight according to risk

➢ Multiply combined axes score with risk score to calculate the **risk weighted combined axes score**.

$$Combined\ score\ x\ risk\ score\ =\ risk\ weighted\ axes\ combined\ score$$

★ Record your 'risk weighted axes combined score' in the box below

Risk weighted axes combined score = _____

Step 8: Compare opacity score to risk weighted combined axes score

➢ Compare **opacity score to risk weighted combined axes score**

$Opacity\ score: risk\ weighted\ combined\ score$

★ Record your 'opacity score' and 'risk weighted combined axes score' in the box below (see Steps 3 and 6 for each score)

Opacity to risk weighted combined axes score = _____:_____

Interpretation of scores

At this stage you should have various scores: an opacity score, scores for each axis, a combined axes score, a risk weighted combined axes score, and an opacity to risk weighted combined axis score.

The following provides a commentary on how to interpret each of these scores and possible actions developers might take in response to such scores.

Part 1: Interpretation of scores

- ★ Insert the applicable scores below to consider your system and its relation to interpretability

a) Interpretation of opacity score

		Opacity	
1	**2**	**3**	**4**
Interpretable			**Uninterpretable**

To consider your model's opacity score:

- ➢ IF the model is HIGHLY INTERPRETABLE (1), the framework should help you reflect upon what work interpretability is doing and the importance of interpretability
- ➢ IF the model is HIGHLY UNINTERPRETABLE (4), the framework should help you reflect upon the measures you may wish to take to control for a lack of interpretability, consider measures to make the model more interpretable, or rethink the design of your model

Comments on score:

b) Interpretation of automation score

Automation			
1	**2**	**3**	**4**
Primarily human determined/action			**Primarily machine determined/actioned**

To consider your system's automation score:

Note that it is not necessarily a negative or positive trait for a machine learning model to be highly automated or be highly human determined. Depending on the intended use of the system, there are reasonable justifications for both. For example, closed-loop blood glucose monitors are well-understood and relatively complete, thereby being good candidates for high automation.

- ➢ IF the model is HIGHLY AUTOMATED (4), this should make you consider seriously whether interpretability might be necessary or desirable, or contemplate whether there should be more human input to the decision/task in question. For instance, in general, when assessing safety and effectiveness of machine learning systems, regulators may need more explanation, more interpretability if the system is highly automated. Where there is no human in the loop, there is often not a good opportunity to rely on the expertise of users to check or interpret the system's output.
- ➢ IF the model has FEW AUTOMATED ELEMENTS (1), interpretability may be more a question of how you deliver, contextualise, and visualise the model's output to allow the user to make a decision or take an action.

Comments on score:

c) Interpretation of adaptivity score

Adaptivity			
1	**2**	**3**	**4**
Static model			**Incremental learning using streaming data**

To consider your model's adaptivity score:

Note that it is not necessarily a negative or positive trait of a machine learning model to be highly adaptive or static. There are risks associated with 'stale' training sets - again, it depends on the intended use. For instance, some applications, for instance, tracking infectious disease via social media may require live data. However, there are also risks associated with highly adaptive models. For instance, neural networks can be vulnerable to 'catastrophic forgetting'

the tendency of a network to abruptly forget previously learned information upon learning new information.[37]

Adaptivity is often a function of engineering choices but also may influence the need for interpretability.

- ➢ IF the model is HIGHLY ADAPTIVE (4), consider carefully whether interpretability may be necessary or desirable. For instance, if the model retrains, this may disrupt the mental picture of the model users have in mind. Consequently, interpretability may assist with contextualising this change. For example, consider a model that does not function well in geriatric populations. Suppose a healthcare professional observes this and mentally adjusts, changing how they use the system. Suppose further that the model is subsequently retrained to better include this population. In this case, interpretability may be important to allow the identification of such changes without solely relying on the notes released with new versions or patches (release notes).
- ➢ IF the model is STATIC/LOCKED (1), interpretability may still be necessary or desirable to contextualise the output of the model.

Comments on score:

d) Interpretation of completeness score

Incompleteness			
1	**2**	**3**	**4**
Comprehensive feature set			**Few relevant features quantified**

To consider your model's completeness score:

Note that complete models are not necessarily accurate models. Indeed, the risk of overfitting may require developers to regularise features. In this way, completeness is listed, not as a desirable trait, but as a trait that leaves less room for interpretability. That is, the more complete a model, the less room there is for interpretability.[38] As noted, the ID Framework does not seek to list desirable or undesirable traits of machine learning but consider attributes of machine learning systems and their relation to interpretability.

As described earlier, completeness works in concert with ground truth. We can only understand if a problem is relatively complete if we know how the features included relate to the decision or task at hand. For example, we cannot know whether a diagnostic tool is relatively complete if we do not have scientific or clinical evidence establishing what features are diagnostically relevant.

It may also be helpful for developers to understand not only whether their model is relatively incomplete but also why. For instance, are there features identified in the scientific and clinical literature as being relevant which are simply not included? Alternatively, is there relatively

little scientific and clinical evidence available to evidence the task or decision at hand, meaning that the incompleteness of the model represents the current dearth of literature?

- ➢ IF the model is HIGHLY INCOMPLETE (4), interpretability may be important as its outputs likely require contextualisation to be clinically relevant or actionable. For example, a triage tool based solely on vital signs would benefit from interpretability so that users can incorporate the model's output alongside their clinical judgment.
- ➢ IF the model incorporates a COMPREHENSIVE FEATURE SET (1), interpretability may be more a question of usability. For instance, where the feature set is highly complete, it may be sufficient to rely more on the predictive accuracy of the model instead of interpretability to establish safety and effectiveness.

Comments on score:

e) Interpretation of ground truth score

Ground truth (lack of)			
1	**2**	**3**	**4**
Robust ground truth			**Contrary to ground truth**

To consider your ground truth score:

Note that ground truth works in tandem with opacity and incompleteness. For instance, if a model is opaque (we do not know what features the model finds significant), it is difficult to link scientific and clinical evidence (ground truth) to evidence the model. Indeed, while ground truth may evidence the inclusion of features in a model, if the model is opaque, it is unclear whether the model relies on these features in line with established ground truth. For example, the model may rely on a feature that is not directly clinically relevant, such as the word 'portable' in x-ray images.[39] Further, to know whether a model is relatively complete, we must know what a complete model would look like – have evidence establishing what features are relevant to the task or decision at hand.

- ➢ IF the model is supported by ROBUST GROUND TRUTH (1), then your feature set is well-evidenced. However, this does not mean that you include all of the relevant features to make that decision or perform that task (completeness) or that you know that the model relies on those features according to your ground truth (interpretability). Having robust ground truth may be necessary to properly contextualise an explanation to a user.
- ➢ IF the model relies on features CONTRARY TO GROUND TRUTH (4) or LACKS GROUND TRUTH (3), you may find it difficult to satisfactorily contextualise the model's outputs. For instance, it may be unsatisfying to tell a user that the model found certain features significant without indicating why these features are clinically relevant or supported in the scientific literature.

Comments on score:

Part 2: Interpretation of combined axes score

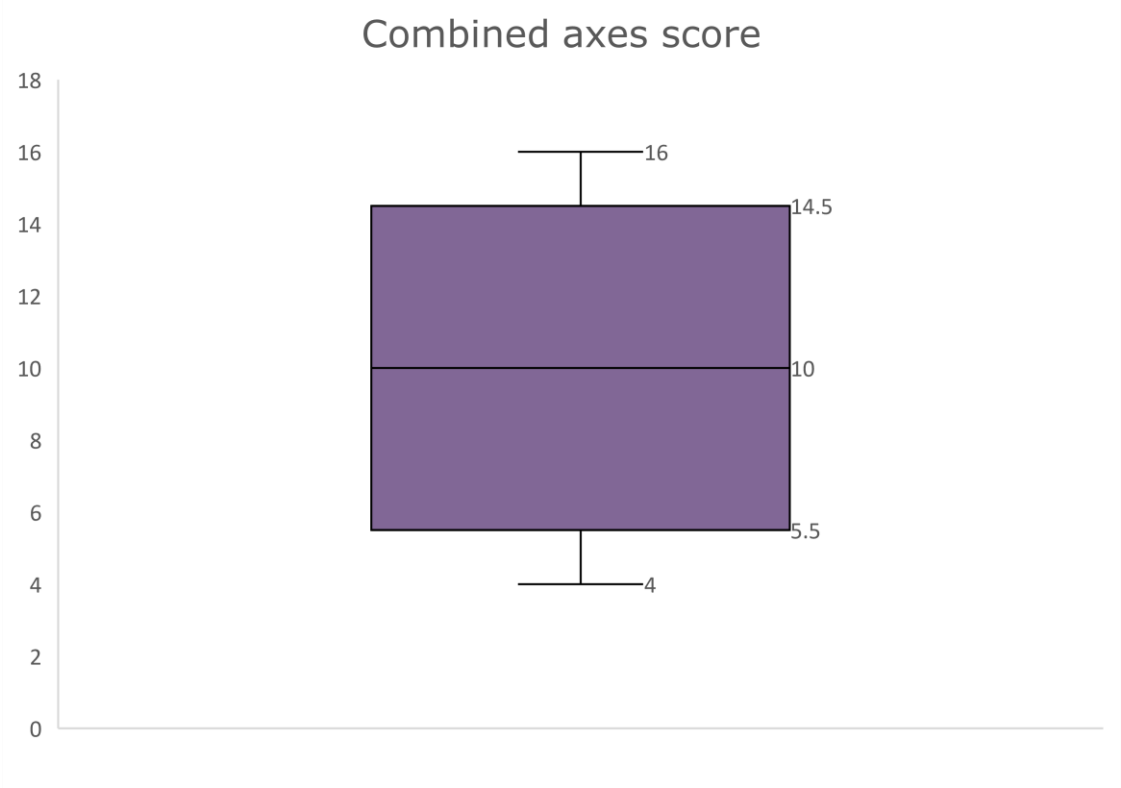
To interpret your combined axes score:

- IF your COMBINED AXES SCORE is HIGH (MAX 16), interpretability of your model may be of great assistance. However, if the model is LOW RISK, this may mean that interpretability is not critical to ensure the device is safe.
- IF your COMBINED AXES SCORE is LOW (MIN 4), interpretability may still be of assistance, especially if your device has a high risk score.

The table below provides an example distribution of risk weighted combined axes scores, the scoring remaining the same but the quartiles (Q1-Q2) changing in response to the distribution of systems.

Score 4-5.5	Score 5.5-10	Score 10-14.5	Score 14.5-16
Q1 LOW	Q2 MEDIUM-LOW	Q3 MEDIUM-HIGH	Q4 HIGH

The box and whisker graph below shows an example distribution of the combined axes scores.



Comments on score:

Part 3: Interpretation of risk score

Risk			
1	**2**	**3**	**4**
No hazardous situation			**Risk to public health**

To consider your risk score:

Risk is to be considered, at least initially, apart from any benefit the system might pose.

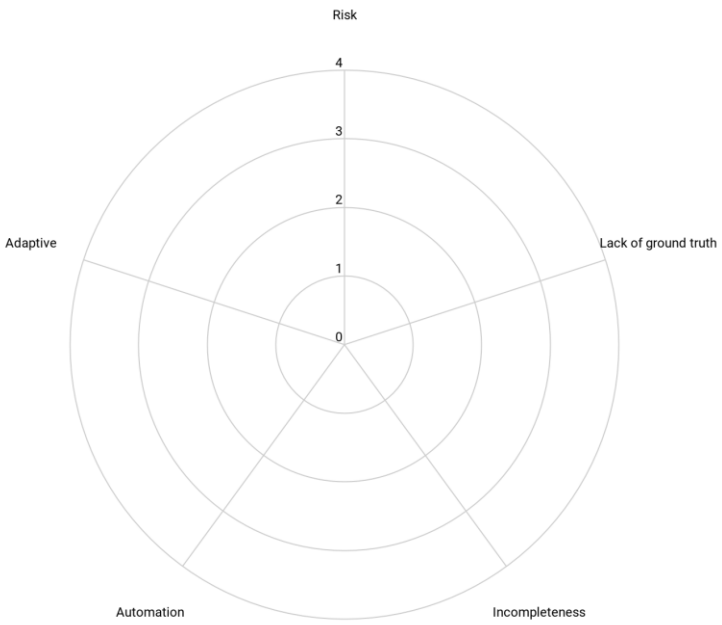
Benefit may be counterpoised with risk. Indeed, as a part of ISO 14971, manufacturers may have to undergo a benefit-risk analysis to consider whether the benefits outweigh the risks.[40] However, the score as recorded does not represent this balancing act. Accordingly, there may be systems that meet a critical need, regardless of the risk and opacity of those systems. Again, the ID Framework does not seek to list desirable or undesirable traits of machine learning but to consider attributes of machine learning systems and their relation to interpretability. However, in rare cases, machine learning systems fulfil such an urgent need that there may be justification for suspending the process of considering interpretability's place in design.

- ➢ IF your system poses a risk of SERIOUS INJURY or RISK TO PUBLIC HEALTH (3-4), consider carefully whether some form of interpretability may assist in reducing these risks. Alternatively, you might also examine other strategies apart from interpretability to minimise these risks, for instance, monitors and alarms.
- ➢ IF your system scores NO HAZARDOUS SITUATION or only poses NON-SERIOUS INJURY risks, interpretability may still assist in reducing these risks 'as far as possible.'[41] Moreover, alternatively, interpretability may also assist in ensuring the system is usable.

Part 4: Plotting scores

- ★ Plot your scores on Figure 1 below to consider the scores of each axis and the relationship between the axes and their scores

Figure 1



Comments on distribution of scores:

Part 5: Interpretation of risk weighted combined axes score

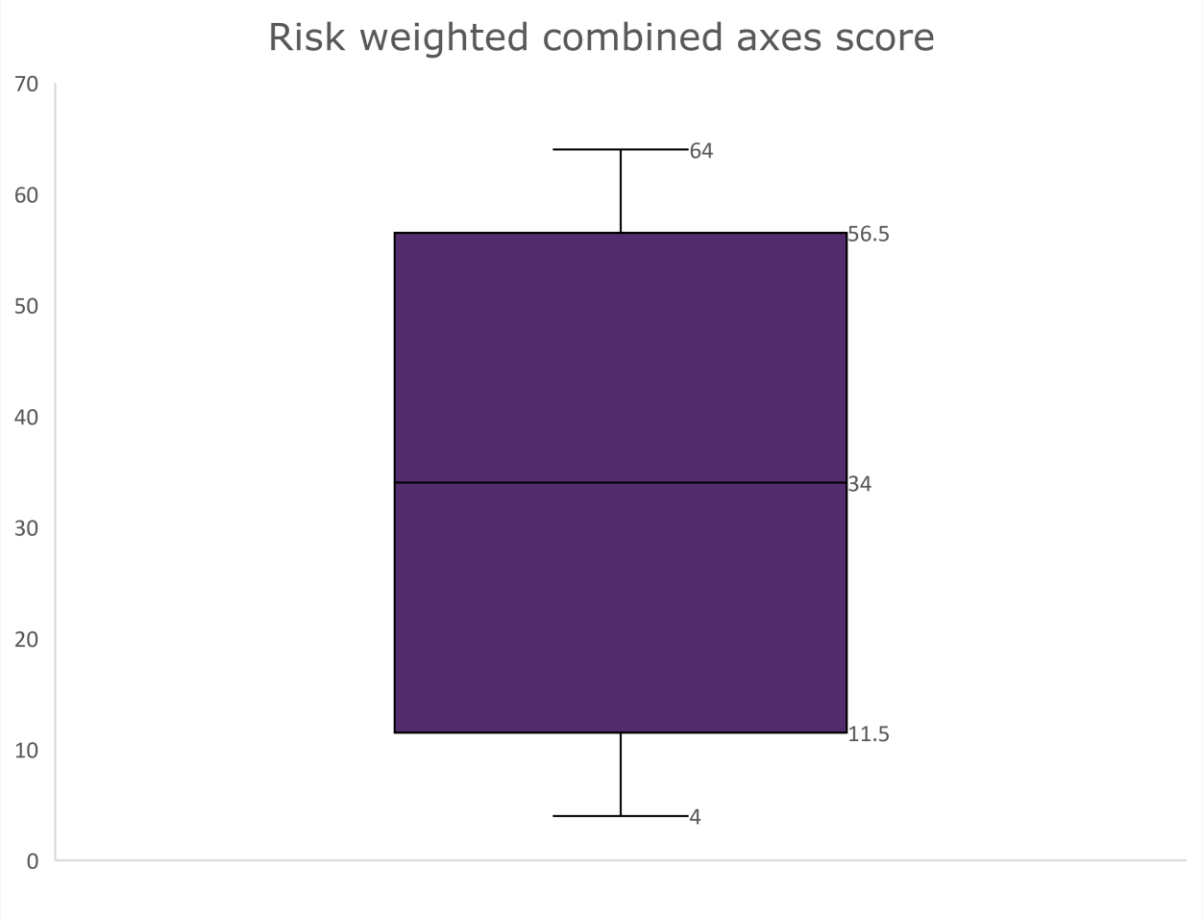
To consider your risk-weighted combined axes score:

- Scores in the range (56.5-64) are considered HIGH, scores in the range (34-56.5) are considered MEDIUM-HIGH. Developers of systems in these ranges should carefully consider their opacity:risk-weighted combined axes score. Moreover, understanding what drove this high score is also important. For instance, was the combined axes score relatively low but the riskiness of the system high? If so, interpretability may primarily be used to bolster otherwise strong ground truth, high completeness, low adaptivity, and low automation scores.
- Scores in the range (4-11.5) are considered LOW, scores in the range (11.5-34) are considered MEDIUM-LOW. Developers of systems in these ranges should consider if interpretability might still assist with the usability of their system. Moreover, understanding what drove this low score may also be important. For instance, was the combined axes score relatively high but the riskiness of the system comparatively low? In this case, while the device may not be a safety concern, interpretability may ameliorate issues relating to automation, adaptability, incompleteness, and lack of ground truth.

The table below provides an example distribution of risk weighted combined axes scores, the scoring remaining the same but the quartiles (Q1-Q2) changing in response to the distribution of systems.

Score 4-11.5	Score 11.5-34	Score 34-56.5	Score 56.5-64
Q1 LOW	Q2 MEDIUM-LOW	Q3 MEDIUM-HIGH	Q4 HIGH

The box and whisker graph below shows an example distribution of the risk weighted combined axes scores.



Comments on score:

Part 6: Interpretation of opacity to risk weighted combined axes score

Opacity to risk weighted combined axes score = _____ :_____

To consider your opacity to risk-weighted combined axes score:

- ➢ Scores that have a HIGH OPACITY and HIGH RISK-WEIGHTED score are systems of most concern. Systems such as these should seriously consider mitigation strategies in regards to each axis score and risk or, alternatively, consider reducing the opacity of their model.
- ➢ Scores that have a LOW OPACITY and HIGH RISK-WEIGHTED score are systems that are highly interpretable, but where interpretability is likely critical to mitigate the risk the system poses and the attributes of the system that might add to the need for interpretability.
- ➢ Scores that have HIGH OPACITY and LOW RISK-WEIGHTED score are systems where interpretability may not be a pressing concern to demonstrate the safety and effectiveness of the system. However, systems such as these should consider interpretability to improve usability of their systems.
- ➢ Scores that have LOW OPACITY and LOW RISK-WEIGHTED score are systems that are highly interpretable and, in any case, do not pose serious risk or have attributes that might create demand for interpretability. Nevertheless, interpretability may still be useful to ensure the system is usable.

Comments on score:

Part 7: Follow up actions and mitigation

- ★ Given the scores entered, interpretation of these scores, and discussion amongst the team, consider the actions and mitigation that result in the box below

Follow up actions and mitigation:

Further reading

For more information, see the report that underpins this tool *Black Box Medicine and Transparency*.

Key resources instructive when thinking through interpretability:

- Liu et al, *How to Read Articles That Use Machine Learning*, JAMA
- Doshi-Velez et al, *Towards a Rigorous Science of Interpretable Machine Learning*, arXiv
- Guidotti et al, *A Survey of Methods for Explaining Black Box Models*, ACM Computing Surveys
- Rudin, *Please Stop Explaining Black Box Models for High Stakes Decisions*, Nature Machine Learning
- Altmann et al, *Limitations of Interpretable Machine Learning Methods*

References

[1] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*. 2017; 1702.08608: 2.

[2] Adapted from: Giorgio J, Landau S, Jagust W, et al. Modelling prognostic trajectories of cognitive decline due to Alzheimer's disease. 2020 Preprint.

[3] Rudin C. Please Stop Explaining Black Box Models for High-Stakes Decisions. *NIPS 2018*. 2018; 2-3.

[4] British Standards Institution (BSI), BS EN 62366-1:2015. *Application of usability engineering to medical devices*. Switzerland: BSI; 2015, 3.16.

[5] Ibid.

[6] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*. 2017; 1702.08608: 3.

[7] Ibid.

[8] Ng A. December 4, 2019. *The Batch*. Deeplearning.ai. 2019.

[9] Vellido A, Lisboa PJG. Handling outliers in brain tumours MRS data analysis through robust topographic mapping. *Computers and Biology in Medicine*. 2006; 36(10): 1049-1063.

[10] Lisboa PJG, Interpretability in Machine Learning – Principles and Practice. In: Goebel R, Tanaka Y, Wahlster W. (eds.) *Fuzz Logic and Applications*. 2013; 15.

[11] International Organization for Standardization (ISO). ISO 9000:2015. *Quality Management Systems: Fundamentals and Vocabulary*. Switzerland: ISO; 2015: 3.8.12.

[12] Ibid.

[13] Liu L, Chen PHC, Krause J, et al. How to Read Articles that Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*. 2019; 322(18): 1808.

[14] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 13th February 2020].

[15] Flach P. *Machine Learning: The Art and Science of Algorithms that Make sense of Data*. Cambridge: Cambridge University Press; 2012: 13.

[16] Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices [2017] OJ L117/1.

Regulation (EU) 2017/746 of the European Parliament and of the Council on *in vitro* diagnostic medical devices [2017] OJ L117/176.

[17] Zech JR, Badgeley MA, Liu M, et al. Confounding variables can degrade generalization performance of radiological deep learning models. *PLoS Medicine*. 2015; 1-15.

[18] Friedman JH, Popescu BE. Predictive learning via rule ensembles. *The Annals of Applied Statistics*. 2008; 2(3): 916-54.

Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices [2017] OJ L117/1, art 61.

[19] National Institute for Health and Care Excellence (NICE), *Evidence standards framework for digital health technologies*. 2019. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies [Accessed 13th February 2020].

[20] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/pdp.html [Accessed 13th February 2020].

[21] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*. 2017; 1702.08608: 2.

[22] Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*. 2016; 3(1): 3-5.

[23] Guidotti R, Monreale A, Ruggieri S, et al. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*. 2018; 51(5); 6.

[24] Ibid.

[25] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*. 2017; 1702.08608: 4-9.

[26] Ordish J, Murfet H, Hall A. *Algorithms as medical devices*. PHG Foundation. 2019: 30.

[27] Gepperth A, Hammer B. Incremental learning algorithms and applications. *European Symposium on Artificial Neural Networks*. 2016: 1-12.

[28] Ibid.

[29] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*. 2017; 1702.08608: 3-4.

[30] Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices [2017] OJ L117/1, article 61(3)(a).

[31] British Standards Institution (BSI). BS EN 62304:2006+A1:2015. *Medical device software – Software life-cycle processes*. London: BSI; 2015: 3.35.

[32] Ibid, 3.9.

[33] Ibid, 3.8.

[34] Ibid, 3.23.

[35] British Standards Institution (BSI). BS EN ISO 14971:2019. *Medical devices – Application of risk management to medical devices*. London: BSI; 2019: 8.

[36] Ibid, 3.17.

[37] Kirkpartrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*. 2017; 114(13): 3521.

[38] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint*. 2017; 1702.08608: 3-4.

[39] Zech JR, Badgeley MA, Liu M, et al. Confounding variables can degrade generalization performance of radiological deep learning models. *PLoS Medicine*. 2015; 1-15.

[40] British Standards Institution (BSI). BS EN ISO 14971:2019. *Medical devices – Application of risk management to medical devices*. London: BSI; 2019: 7.4.

[41] Ibid, 4.2.

The Black box medicine and transparency report was funded by the Wellcome Trust as part of the 2018 Seed Awards in Humanities and Social Sciences [Grant Number: 213623/Z/18/Z].

We thank the Wellcome Trust for their support.

The PHG Foundation is a non-profit think tank with a special focus on how genomics and other emerging health technologies can provide more effective, personalised healthcare and deliver improvements in health for patients and citizens.

For more information contact:

intelligence@phgfoundation.org

UNIVERSITY OF CAMBRIDGE

phg foundation

making science work for health