Building Explainability and Trust for AI in Healthcare

December 2019

XAVIER HEALTH

TABLE OF CONTENTS

Acknowledgments

This paper was developed under the leadership of the Xavier Health program at Xavier University in partnership with industry professionals, as a planned output from 2018 AI Summit. We would like to thank everyone who contributed to the creation and the review of this paper – without their work, this paper would not have been possible[1]. We also want to acknowledge the significant contributions from the following people:

Eileen Alexander, Xavier University

Bob Banta, Eli Lilly

Tara Feuerstein, Farm Design

Steve Frigon, The Christ Hospital

Lacey Harbour, Ken Block Consulting

Roger Hecker, MobileODT

Dr. Christian Johner, Johner Institute

Tripti Kataria, MD, MPH, FASA

Eileen Koski, IBM

Bob Kruth, J&J

Cindi Linville, Best Sanitizers

Zinatara Manji, GSK Consumer Healthcare

Mac McKeen, Boston Scientific

Robert Z. Phillips, Siemens Healthineers

Sundar Selvatharasu, Sierra Labs

Dylan Sinks, Lilly

Sunnie Southern, Viable Synergy

Scott Theil, Navigant

Danny Tobey, DLA Piper

Sylvia Trujillo, AMA

Camille Vidal, GE

Paul Westfall, AMA

May Yamada-Lifton, SAS

Jeremy Zhang, Abbott

David Zinger, Medela

Our hope is that this paper provides the foundation for new learnings and best practices in this rapidly evolving field to help deliver the promise and potential of AI.

Pat Baird, Philips

Rohit Nayak, Electronic Registry Systems

Kelly Nienburg, PwC

[1] Please note that the opinions and viewpoints expressed by the contributors do not necessarily reflect the opinions and viewpoints of their organizations.

Introduction

Background

Over the past few years, Artificial Intelligence (AI), and more specifically Machine Learning (ML) technology has experienced rapid adoption in the healthcare space as tools for diagnosis and decision-making. Such tools are intended to address challenges in the health care system to both process and apply rapidly proliferating medical findings to practice, as well as to deliver on the promise of personalized and precision medicine. One of the most pressing challenges facing the widespread adoption of AI technology is that it is difficult to comprehend the 'black-box' nature of AI design that is often associated with complex ML algorithms.

To address the potential mistrust and misuse that may come from the perpetuation of the idea that AI technology represents a black box, the explanatory nature of AI design must be addressed with greater urgency. Coinciding with an increase in the public discourse over bias in ML technologies, the notion of AI fairness and avoiding AI bias has received increasing attention in academic circles. Failing to address these concerns may ultimately lead to the rejection or slow the adoption of AI technologies as potentially biased, lacking sufficient specificity and/or sensitivity, too difficult to explain and/or understand, and therefore not 'trustworthy' enough to use in healthcare applications.

This urgency has been further heightened by the "Explainability Requirement" within the EU General Data Protection Regulation (GDPR) which went into effect in 2018. Specifically, GDPR Articles 13-15 and 21-22 outline requirements related to the basic concept that when a decision is generated solely from automated processing (no human intervention), including profiling, the data subject has the right to receive an explanation of how the decision was rendered. This "right to explanation" has AI experts debating the extent to which this is applicable, particularly as it relates to healthcare applications.

One recent example of an initiative to drive toward explainable AI applications is the Explainable Artificial Intelligence (XAI) initiative from the Defense Advanced Research Projects Agency (DARPA)[2]. The XAI initiative aims to deliver a toolkit library of ML and software engineering tools that can be used for the development of explainable AI models[3]. Their hope is to usher in the third-wave AI solutions which can understand the context in which they function and can characterize the real-world phenomena. Its main aim is to:

- Explain the decisions and processes of the AI application
- Understand model strengths and weaknesses
- Predict how the system may behave in the future

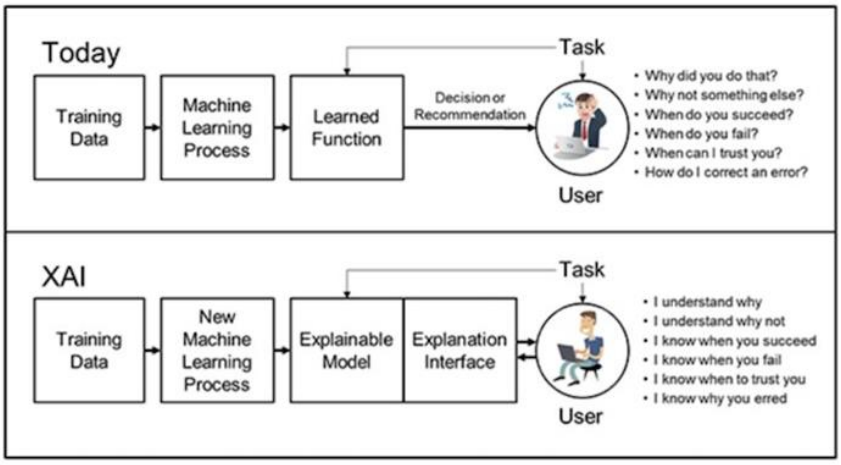

Figure 1 - XAI; Source: DARPA3

[2] https://www.darpa.mil/program/explainable-artificial-intelligence
[3] Gunning, David. "Explainable Artificial Intelligence." 2017. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

To address the gap between the black-box nature of many AI systems and the necessity of increasing trust in machine learning technology, this paper focuses on the discussion of AI for healthcare related applications, describes approaches to revealing the inner workings of AI applications.

The notion of Explainability is explored across a few key systems development characteristics; these include Presentation, Adoption, Design, Knowledge, Performance and Updates. These characteristics will be reviewed across a few common stakeholder segments. In addition, domain knowledge related to the usage of data will receive special attention due to the unique sensitivity of healthcare application-related data in the regulatory sphere.

Overall Goals and Objectives

This paper is intended to open a discussion of current approaches for Explainability and trust by highlighting commonly used systems-development constructs oriented towards healthcare. By highlighting a series of considerations and questions, the reader is enabled to make their own conclusions. Finally, the paper concludes with an in-depth review of how the degree of Risk can drive the level of Explainability necessary for applications of AI within a healthcare setting.

This paper is not intended to be a standard, nor is this paper trying to advocate for one and only one approach to developing Explainable AI applications. This paper is also not intended to evaluate existing or developing regulatory, legal, ethical, or social consequences these approaches. Note that since “cybersecurity and AI” is such a rapidly developing topic, it is not discussed in this paper.

Audience and Stakeholders

The intended audience of this paper is broad, and includes developers, implementers, researchers, quality assurance, regulatory affairs, validation personnel, business managers, regulators and end-users faced with challenge of assessing and building trust in AI healthcare applications.

2. Presentation and Adoption

Presentation
What are the different methods for delivering an output that incorporates Explainability?

For user-facing applications, what approaches might be considered to deliver confidence and transparency in output (e.g., performance metrics, labeling standards, etc.) ?

Adoption
What level of explanation is necessary to drive confidence in the output of the system to different types of stakeholders? What is good enough?

To realize the full benefit these technologies offer, clinicians and patients must understand and accept the validity of the role of the technology as it is interwoven throughout their daily activities. Explainability is essential to engendering trust in information and decisions derived from AI/ML/CLS applications. In the absence of trust, neither physicians nor patients are likely to use such tools when the decision to be made is a consequential one, such as a diagnostic determination or treatment decision. The bar may be lower for a "general wellness" device or app that may not be well validated, but that may only be used for motivational or entertainment purposes to encourage individuals to be aware of their behaviors and activities as it relates to their health and fitness.

This section addresses key "user" considerations related to the "presentation" and "adoption" of AI/ML/CLS applications related to patient diagnostic and treatment as well as healthcare providers adoption of such tools in the practice of medicine and patient care. The Presentation and Adoption domains relate to users' belief and perception of their need, the credibility of the AI application itself, the value of the outcome, and recommendation based on the output by a trusted professional, family member or friend.

Establishing Trust
The need for AI applications is based on the challenge of addressing the complexities posed by the combination of increasing numbers of patients with multiple co-morbidities, highly technical diagnostic modalities (imaging, genomics, etc.), rapid increases in medical literature and variability in practice patterns. Burden of Disease (BOD) and market opportunity can further spur efforts to commercialize products that provide solutions to address these needs. The flow chart below represents the iterative journey from identification of unfulfilled health needs through reimbursement and use by late adopters:

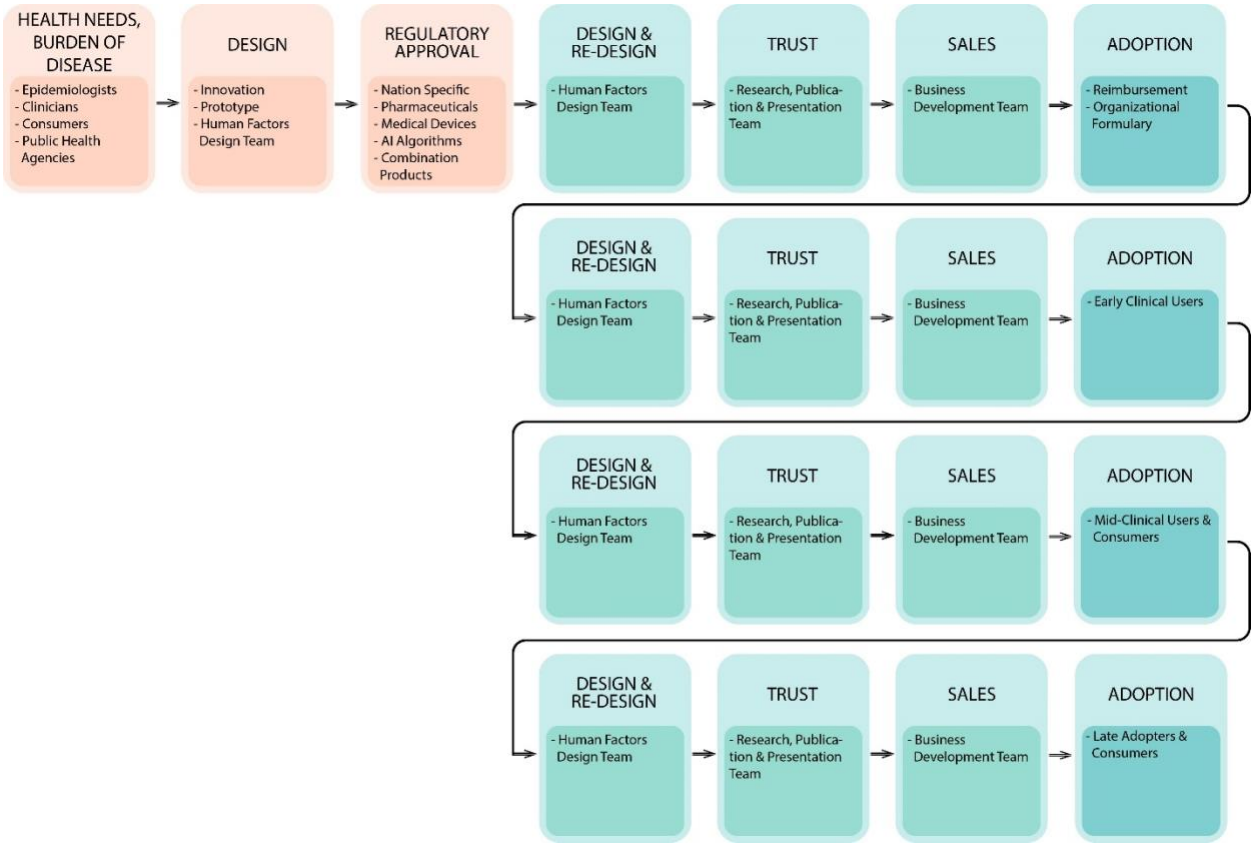

Figure 2: Trust and adoption increases using an iterative series of re-design, scientific publications, presentations, and addressing business development concerns for reimbursement and differing levels of users. (E. S. Alexander & S. L. Alexander, 2019.)

Additionally, there are likely skills and roles such as Health Economics and Data Scientists not emphasized in this flow chart and need additional consideration, particularly as a part of the Design Team.

Increasing levels of trust are built to support reimbursement and adoption by early, mid- and late adopters. At each level, the human factors design (HFD) team addresses Explainability. At each level, performance is analyzed and published in peer-reviewed medical and technical journals. A series of planned publications addresses post-market effectiveness and value. Business development brings new issues back to the HFD team at each iteration, from each successive group, i.e., formulary approvers, clinical users and consumers. By presenting successively and appropriately to early, mid- and late adopters, key opinion leaders and business development professionals increase market credibility, i.e., trust.

The overall success of AI applications in healthcare, medicine, and consumer health and fitness resides on the stakeholders accepting this technology in their everyday lives and in the practice of medicine. Applications and products that contain AI for healthcare purposes should be designed to provide the user with an easy to understand explanation of the required data inputs and the output recommendation of the product. This will promote an understanding and trust of the AI application and promote adoption

Although not specific to AI, the FDA has provided multiple guidance documents related to Digital Health and Mobile Medical Applications (MMA), Software as a Medical Device (SaMD), and Medical Device Data Systems (MDDS). In addition, the International Medical Device Regulators Forum (IMDRF) has provided several definitions in this area and a framework for Digital Health tools that are being applied and integrated into the regulatory process and development of applications and solutions.

The resulting emphasis is on three primary markets for AI healthcare applications and how they are presented to users. These involve the presentation/publication/adoption to end users in:

1. Healthcare system/physician-driven clinical decision support and,
2. Engagement by patients and the acceptance of the AI application in the diagnoses and recommended treatment options,
3. Insurers and reimbursement organizations.

Note that a deeper discussion of this topic can be found in the Appendix.

Key Principles to Follow for Healthcare Applications

Clinicians practice medicine, which is evidence based and relies upon a combination of experience and the scientific method[4]. However, in machine learning, specific data inputs and predictive methodology are obscured. For physicians to trust AI devices/programs, generalized explanations of the methodology, clinical rationale and limitations will be required. The validity of the AI will be judged on how the AI compares to the current best practices and the gold standard through peer reviewed literature. It will be important to explain the clinical testing of AI technologies and the confirmation of the results by physicians.

For clinicians to trust AI healthcare applications, simplified approaches to explanations will be required of the input and the output. This is similar to how we don't understand the details of how a CT or MRI scanner works, we do understand the product of the scan. It is important to explain the appropriate use of the AI, its clinical rationale along with its limitations.

The validity of the AI application will be judged on how the AI application compares to the current best practices, real world evidence, and the gold standard through peer reviewed literature. It will be important to explain the clinical testing of AI applications and the confirmation of the results by clinical experts. In particular, if the AI application offers a recommendation that differs from standard practice, such explanations will be even more critical.

[4] Fogel, Alexander, Kvedar, Joseph, *Artificial intelligence powers digital medicine*, npj Digital Medicine (2018) 1:5; doi: 10.1038/s41746-017-0012-2.

Three key principles apply to the presentation and adoption of healthcare related AI applications as it relates to Explainability and building trust:

1. Adoption of AI application in the practice of medicine will most likely follow the current model of introducing new products and technologies to the market:

   a. Regulatory review and agreement,
   b. Peer to peer sharing/trust, Key Opinion Leader recommendations,
   c. Presentation at meetings/scientific congress,
   d. Peer reviewed Literature, and
   e. Coverage reimbursement (where necessary)

2. Engagement builds familiarity and trust.

   a. Clinicians are presented with the AI applications and through use and experience develop a "trust" in the AI application and then they will adopt it in their practice.
   b. Once the physician adopts the AI application into their practice, they will present it to the patient and instill their "trust" in the output for the benefit of the patient, and the patient will choose to accept (adoption) of the AI.

3. Risk plays a key role in this process. One may initially think that highly complex applications may take longer to adopt than simpler applications, but if the complex application provides greater benefits or less risk than other alternatives, adoption may be accelerated. Therefore, the role of Benefit-Risk analysis must be considered.

For some forms of AI/ML/CLS the adoption path for new AI technology will be neither new nor different. Such AI applications in healthcare settings will travel an already established road. Newer AI systems, such as Adaptive Clinical Decision Support (CDS), that rely on continuous learning from real-time data feeds may pose new challenges to both regulators implementers and will require modifications to current standard practice.

Note that there is an alternative flow–patients may see the technology in news reports or via social media and ask the clinician about the technology and request information from the clinician and potentially ask for that technology to be used in their situation. With this in mind, public acceptance is important and is also dependent on the type of device. For example, a wearable will gain public acceptance when individuals see reported evidence of the benefits in sources they trust. They will accept this technology and adopt it. They will drive the acceptance with their clinician. Conversely, it may be good to note the current state is that most patients trust their physicians to select the right application/tool/device for their treatment, and to also ensure its correct use.

Physician Perspective

One key factor for healthcare that is different from other industries is that clinicians rely on their peers and Key Opinion Leaders (KOLs) when making healthcare choices. To be successful, a product must win over clinicians first as part of the journey and then the process can be translated to the patients[4].

The benefits of the system should be explained in clear language to the end user. From our experience, the concept of AI is "cool" but the AI application should provide a diagnosis in the manner to which the provider is accustomed: the same "lab results" should be presented from the AI device as much as possible. To build confidence in the AI system, when initially rolling out the AI capabilities, the clinician compares his or her own diagnoses with system generated diagnoses. By collecting these data from early adopters, a construct validity database is built, and available for analysis and publication. Construct validity is a standard pathway to test and establish trust in new laboratory and technological methods. After clinical validation and market penetration, we expect late adopters to accept the instrument's diagnosis as they would any diagnosis from a lab.

Excerpt from "Scalable and accurate deep learning with electronic health records"

The use of free text for prediction allows a new level of Explainability of predictions. Clinicians have historically distrusted neural network models because of their opaqueness. We demonstrate how our method can visualize what data the model "looked at" for each individual patient, which can be used by a clinician to determine if a prediction was based on credible facts, and potentially help decide actions. In our case study, the model identified elements of the patient's history and radiology findings to render its prediction, which are critical data points that a clinician would also use. This approach may address concerns that such "black box" methods are untrustworthy.

https://www.nature.com/articles/s41746-018-0029-1

Healthcare is a field with a lot of variation, both over time and by location. There may be situations where a process and its results are radically different from an output perspective, and yet completely and immediately understandable – e.g. qualitative testing replaced by quantitative results or synthesis of large amounts of data from EMR and a new device to reach a conclusion or report indicating the disease state and treatment plan

If the clinician understands that pathway, which is similar to their own information synthesis methodology, they will more readily understand and adopt the new technology.

Patient Perspective

It is common practice for patients to seek information about their health, condition and treatment options on the Internet from the many health-related websites and sources of medical information. More and more patients are armed with information about their health before they engage the healthcare system at large and make their first visit to a doctor or clinic.

Patients will look to clinicians for recommendations to build trust in the AI application and output based on valid scientific evidence. Once the physician accepts the AI application, they present it to the patient (Presentation mode) and communicate the benefits of the technology. This builds trust with the patient for them to accept the AI application (Adoption mode). As we're seeing with numerous consumer-centric digital health applications, patients or consumers are perhaps less likely to question the output of their technology or devices.

As previously noted, patients may also learn of an application from other sources and may approach their clinician with their own ideas – in this case, the patient may already have developed trust in the application before even talking to their clinician.

Presentation Considerations

Some key considerations for enhancing Explainability in AI for the Presentation domain include:

- What types of information do we present to the user in a real-time setting?
- How do we do this without achieving information overload?
- How do we meaningfully explain how the data is collected, used and applied across various situations and workflows?
- How does the data change over time?
- How do we drive trust and confidence in the data sources used to enable the AI capabilities?

Note: A more complete listing is available in the Appendix

Adoption Considerations

Some key considerations for enhancing Explainability in AI for the Adoption domain include:

- Does the technology result in additional steps or tasks that may impact Adoption?
- How does the solution ensure that end users (whether clinicians, patients or caregivers) are educated on the use of their devices, how the data is being collected, and that the language is clear?
- What is the optimal level of training necessary for the end users? Should this be documented/mandated by regulatory bodies?
- What approaches might be considered to deliver confidence and transparency in output?
- What level of Explainability is sufficient? Is it the same for all users?

Note: A more complete listing is available in the Appendix.

3. Design and Knowledge

Design
There are a variety of stakeholders required for a successful design, approval, and use of AI in healthcare. What level of explanation is necessary to give confidence in the design of AI to different types of stakeholders? How can Explainability be built into the design?

Knowledge
What level of data and domain understanding is required to begin an AI project in healthcare?

How do we evaluate if we are using the right data?

What tools/techniques should we use to evaluate data integrity, data completeness, and data bias?

Explaining AI Design to Business and Non-Technical Stakeholders

Business and non-technical stakeholders can include project managers, company executives, Clinical Affairs, Chief Medical Officer, Regulatory Affairs, Quality teams, sales, manufacturing, legal functions as well as external investors, and others. Each stakeholder has different concerns with the trustworthiness of the accuracy of the AI algorithm data output. Project managers need to have enough understanding of the algorithm outputs in order to successfully lead the project to completion. Company executives (such as top management) need to have confidence in the AI algorithm design and development process such that they are prepared to provide adequate resources to meet the commercialization and lifecycle management needs.

In order for business and non-technical stakeholders to have confidence in the artificial intelligence output, key aspects of design must be considered including relevant details around the design of the underlying algorithms, preferably accompanied by a visual representation. The algorithm design can support Explainability by providing an output that gives the user, in terms the user understands, a description of how the algorithm arrived at its conclusion (i.e., relational context, statistical match, confidence, etc.)

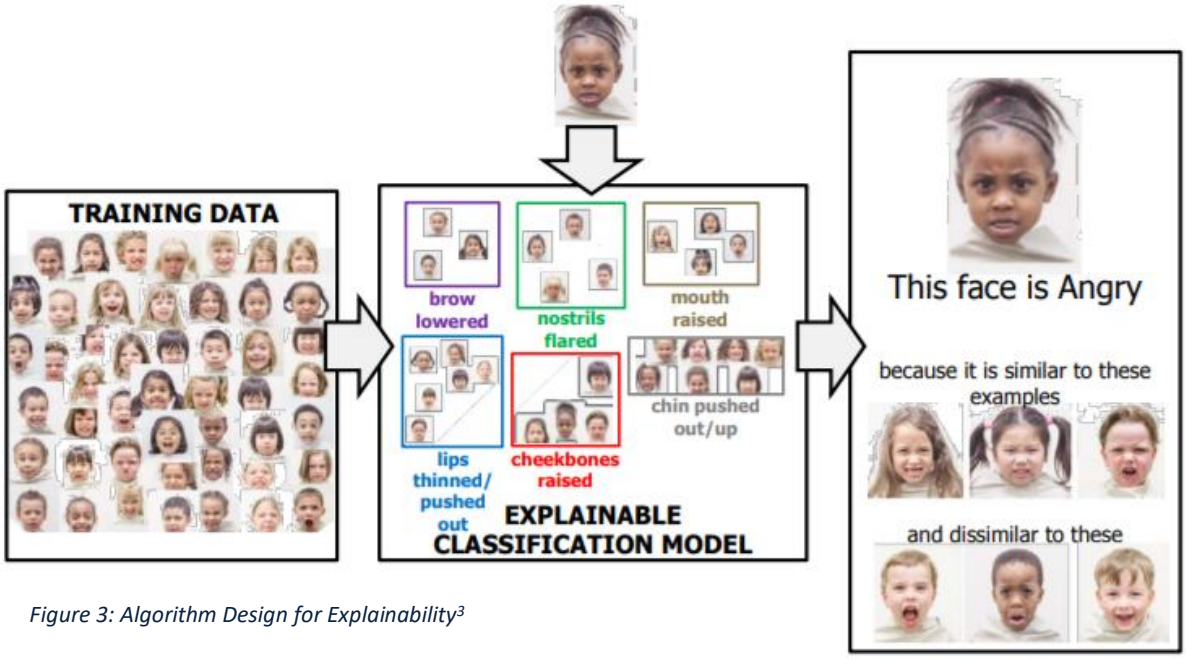

Figure 3: Algorithm Design for Explainability[3]

For example, one approach shown in Figure 3 operates such that once the algorithm arrives at an output, what generated the output is tagged and the tag is shared with the user for interpretation. Since the algorithm is able to provide the user with a reason for the output, the stakeholders are able to gain confidence in the ability for the AI to support the user needs. For systems with multiple algorithms, this process should be repeated for each algorithm.

Additionally, visual representation of AI to the user can provide stakeholders the necessary high-level understanding of how the AI is arriving at an output. Visually representing AI can be accomplished many ways, but a simple to understand diagram will give stakeholders the necessary confidence. AI could be represented via a basic diagram that provides explanation of the input, how the input is analyzed, and the output or decision of the AI. Given a high level visual, users will be able to quickly gain a basic understanding of the AI decision. The important questions around how "basic understanding" is defined and whether it is sufficient will inevitably likely evolve with technology advances and related practitioner and industry experience.

Explaining AI Design to Decision-makers

Decision-makers involved in the usage of AI applications in healthcare broadly include both healthcare providers, insurers, and consumers, particularly if the application does not require the involvement of a healthcare professional. Clinicians rely on mechanistic explanations that correlate observations (inputs) with results (outputs). They use patient history and physical exams, laboratory analysis, radiological imaging and evidence based critical thinking to deduce the cause of a condition and make a clinical decision. The method in which a physician makes decisions is critical to formulating explanations of AI applications to clinicians.

To a physician inexperienced with AI technologies, implicit trust in black-box models may need to be a starting point where a transparent input points to an output derived by an AI algorithm. While certain aspects of this circumstance are not unusual due to current reliance on FDA regulatory review, approval and oversight processes, ultimately, the black

box needs to be explained in a logical fashion with clear supporting evidence of its conclusions that mirrors the pathway a physician might normally follow to make clinical decisions.

The premise of decision making by AI is the first hurdle. The physician will want to understand in broad terms the logic of the algorithm design. In other words, what is the "critical thinking" of the model? One potentially easy method of explaining AI decision-making lies in logical visual flow charts that apply similarly across both decision-making stakeholders as well as business stakeholders. Flow charts help provide concrete evidence of existing process flows of otherwise opaque black-box technologies.

Decision-makers also want to know how to evaluate trust within the AI application, essentially asking the question "How do you know your AI model works?" Information such as the origins of the training data set, the representativeness of the population, the inclusiveness of data types, the training process, and how inputs and outputs are paired represent important factors to consider. Model validation and the source of test data will also need to be explained. This would be an opportune time to reveal the thresholds and criteria for predictive results and how it would impact its effects on the patient and how the results of the model compare to the current gold standard of diagnosis.

The method of delivering the system output to the physician is also a key aspect of decision making for a physician. Clinicians already suffer from alert and information fatigue and are inundated with a large amount of paperwork. The method of delivery should be one that indicates a level of urgency for intervention if necessary, and yet does not over escalate urgency when not indicated.

If the model is one that is fully prescriptive or automated without input from the physician prior to intervention, the physician must understand that aspect of the algorithm along with the thresholds of the decision making. Fully prescriptive or automated devices/algorithms without physician input carries a level of risk to the patient due to lack of or limited intervention by a human. This type of device would be held to a higher standard of accuracy and scrutiny as a result.

Regulatory Considerations of AI Design

Regulating AI technology in the healthcare industry represents a unique challenge to regulators in this space. Relevant requirements and guidance include:

- FDA's Quality System Regulations (21 CFR part 820)
- Several FDA Guidance Documents e.g. General Principles of Software Validation
- European Medical Device Regulation MDR
- ISO 14971:2019 "Medical devices – Application of risk management to medical devices"
- IEC 62304:2006 + A1:2015 "Medical device software – Software life cycle processes"
- ISO 13485:2016 "Medical devices — Quality management systems — Requirements for regulatory purposes"
- IEC 62366-1:2015 "Medical devices – Part 1: Application of usability engineering to medical devices"
- European General Data Protection Regulation GDPR

According to these requirements, manufacturers must satisfy several different design parameters which may have an impact on the Explainability of the AI device. Manufacturers must define and follow software life cycles processes and compile a use specification including intended medical indications, intended patient population, intended use environment and intended user profile. Manufacturers must specify the stakeholder and human factors related requirements (this includes user requirements, for example by healthcare professionals or patients) and specify the design inputs and outputs, including the user interface specification.

Manufacturers must protect the privacy of healthcare data and give "data subjects" the right not to be subject to a decision based solely on automated processing which significantly affects them. In addition, manufacturers must perform state-of-the-art verification and validation of the product and ensure the repeatability, reliability, and performance of the device. Risk mitigation and controls must be in place including cost-benefit analysis, residual risk mitigation, and control usability related risks and users need to be supplied with clear instructions for use that reveal any residual risk.

If a clinical evaluation is necessary, they need to consider the applied software algorithms to prove the technical equivalence of the AI device with predicate devices at a minimum. Specific functional and performance requirements of off-the-shelf software (OTS) components need to be verified and meet requirements. All computerized software systems need to be validated when used in a quality management system. Competencies of individuals involved in development of the AI product need to be specified.

Explainability for Technical Stakeholders

AI developers need to consider several key inputs when developing AI applications to meet regulatory requirements that are closely related to the technical performance of the device. These include but are not limited to:

- Specification of user profiles, use environment or already specific user interface requirements
- Ground truth respectively gold standard to compare with or alternatively dedicated requirements such as targets. For example, for specificity, sensitivity, ROC curves, F1 scores, accuracy, costs of errors e.g. false positives etc.
- Risk (probability and severity of harm) if the specifications are not met
- Data characteristics to be dealt with (e.g. ranges, variations, outliers, missing values)
- Performance requirements (e.g. response rates, data volumes to be handled, etc.)
- Frequency of update of the model (e.g. trained once? Or updated continuously? Or updated periodically?)
- Run-time environment (e.g. smart phone versus cloud service)

AI Designers should consider compiling the following documentation in support of Explainability to technical stakeholders listed in the following table:

DOCUMENTS[A]	RELATIONSHIP TO INTERPRETABILITY AND EXPLAINABILITY
User interface specification	Specification of how the system • shows the results • explains the results • informs the users about limitations and preconditions • warns the user on wrong input, if preconditions are not met and if results are not trustworthy
Test documentation (e.g. test plans and results) for product, software, components	• System shows and explains the results (as specified) • System shows warnings if preconditions are not met (as specified) • Description of how the system's performance was tested[B]
Human factors and usability tests with intended users and intended use environment	Evidence that users understand • the results (e.g. treatment recommendations, diagnosis) • the limitations of the systems • whether preconditions are met • how trustworthy the results are
Online help, instructions for use, training materials	• the limitations of the systems • the preconditions the system relies on • the residual risks • how the algorithms work, on which data it has been trained
Clinical evaluation, risk management file	• Rationale why model (type, hyperparameter etc.) is the most suitable one e.g. a benchmark that compares the performance of different models • Rationale why model produces results that are reliable, repeatable, reasonable (e.g. by white-box surrogate, intermediate outputs, description of training data etc.) • Explanation of limitations and residual risks (this also might require a white-box description or intermediate outputs)[C] • Evidence and validation that the results / output of the system facilitates the intended medical purpose e.g. better diagnosis, treatment etc.
Software development plan	• Specification of competencies e.g. of data scientists, human factors experts, medical experts • Methods for verification and validation e.g. of usability • How will the model be updated?
Model description	• What type of model was used • The limitations of the model (e.g. linear regression can't handle interactions automatically)

	• How well the model can be explained
Data description	• The data the model has been trained on • Potential biases / problems / limitations of the data • How well will training data match with application population?

[A] Only documents and other artifacts are mentioned that are relevant to the Explainability and interpretability

[B] If the wrong scheme was used (e.g. model is tested on the same data as it was trained on), the performance measures will be severely biased. And it's easy to make mistakes when measuring the general performance

[C] The limitations and risks of the training data needs to be considered. If in the data there were only women, the medical device should be only applicable to women. Or if the variance of the outcome was high for younger people, then there is a higher risk that predicted outcomes for young people might be wrong.

Knowledge of Data in AI Technology for Healthcare Systems

AI applications rely on input data to learn, develop, and act. Ensuring a high-level of data confidence and data quality is important to the training and performance of these systems. Therefore, it is critical to maintain careful documentation, disclosure, and accounting of the following aspects of datasets used for AI training: the origin or source of the data, the data collection process, the data model, and methods for data curation. In addition, processes that measure, enhance, and monitor data integrity as well as data usability should be transparent. For example, mitigation of data quality issues and processes that address the lack of standardization, inconsistent data formats, and errors in the dataset should be documented.

Data sources are varied in origin and include everything from unstructured information, such as conversations on social media, to automated physiologic monitoring caches and data gathered by organizations that adhere to standards that support semantic interoperability that are made publicly available. While it may be difficult to document data standards for a variety of data sources, even where data has its origins in organizations with standards-driven data capture, varied data models and formats may impact data accuracy and consistency.

It is especially critical when evaluating data sources for use in continuous learning systems to consult a variety of subject matters experts such as data scientists, statisticians, epidemiologists, engineers, clinicians, and regulatory experts to ensure that there is no bias within the process of selecting input data. These experts must analyze the data collection, curation, and maintenance techniques and ensure that data is generated using valid scientific and regulatory techniques. These experts should also be involved in training and monitoring the continuous learning systems.[5] In addition to the evaluation of input data, the integrity of data treatment during its lifecycle within the continuously learning system must also be documented. The relationship between the input data and the accuracy, consistency, and validity of the AI output should be documented.

The development of AI applications holds great promise but growing concerns are being raised surrounding data and algorithmic bias in AI applications in a variety of applications including employment, banking, and criminal justice.[6] Bias

[5] Xavier Health, Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare (Aug. 2018), https://www.advamed.org/sites/default/files/resource/perspectives_and_good_practices_for_ai_and_continuous_learning_systems_in_healthcare.pdf.

[6] AINow, Algorithmic Accountability Policy Toolkit, (Oct. 2018), https://ainowinstitute.org/aap-toolkit.pdf.

can take many forms including unintended bias in the data used or derived, and can occur in multiple forms. Some examples include sample bias, algorithm bias, assessment bias, prejudicial bias and even algorithm bias.

Data Integrity Considerations for AI Knowledge

Human intervention within the AI decision making process can introduce unintended bias, so data integrity should be built within the system programmatically to ensure continuous monitoring. Data must be collected, prepared, analyzed and applied with discipline and consistency to ensure the continual integrity of the system. There must be reasonable clarity into the types of data from which the system learned, and consumers must understand contextually (functionality and flexibility of the model) how the model was qualified.

Documentation regarding characteristics of datasets should be accompanied by an assessment of potential bias within the input data set, bias in selection of algorithms, and potential bias from general human interaction and any efforts made to mitigate these risks. Accounting for unintended bias in data sets and algorithms is a central metric of data quality and a key to acquiring trust in the AI application from internal and external stakeholders as well as regulators.

Moreover, the industry should look to develop standards regarding avoidance of bias in AI technology. For example, IEEE is currently developing a standard to "help users certify how they worked to address and eliminate issues of negative bias in creation of their algorithms."[7] IEEE also lays out potential elements to include in a certification that could also be used as part of an assessment of potential biases including procedures and criteria in how the validation data set was selected for bias quality control, setting boundaries on the uses of the algorithm to help prevent unintended consequences, and contemplating end user experience and potential incorrect interpretation of an AI systems output.

[7] IEEE, p7003 – Algorithmic Bias Considerations, https://standards.ieee.org/project/7003.html.

4. Performance

Continuous Learning

Pe
Wh

How do updates impact the Explainability of the AI?

How can visibility into new data and updates help enable Explainability?

?

The methods of assessing and explaining performance of Continuous Learning System (CLS) models are largely consistent with predictive modeling methodologies. As such, it is helpful to review performance widely across predictive modeling techniques. A model's performance is largely dependent on the quality and representativeness of the data for the domain and intended use. An assumption for this discussion is that the input data is of a sufficient quality and breadth to build a model acceptable for its intended use and risk.

The first distinction when considering performance is between supervised and unsupervised learning. Unsupervised learning does not have a target or response variable, but rather finds patterns or relationships between the inputs. Common unsupervised learning applications are clustering analyses like hierarchical or k-means. The performance metric most clustering analyses seek to minimize is the within-group sum of squares (WGSS)[8]. The results of unsupervised learning methods are typically assessed qualitatively. The clusters of a segmentation model, for instance, should be reviewed by stakeholders for verifying hypotheses or providing insight for more targeted analyses.

Conversely, supervised learning methods have inputs and a target output variable. The model enumerates the functional relationship between the inputs and outputs. The two main categories for supervised learning are regression (quantitative target variable) and classification (factor or categorical target variable). CLS systems function similarly regardless of whether they are classification or regression, but the metrics used to quantify performance are different.

There are two main dimensions of performance for a supervised model, bias and variance. Bias refers to the accuracy of the model. Bias is measured by first splitting the data set into training and test sets. The model is built or trained using the training set, and then applied to the input variables in the test set to generate a prediction set. The amount the prediction set differs from the actual output values in the test set is the model's bias. The most common metrics for regression bias are Mean Squared Error (MSE) and Mean Average Error (MAE)[9]. The lower the values the better the fit. Other common metrics for assessing performance of regression models relative to each other are adjusted R^2 and AIC (Akaike Information Criterion)[10]. For classification, the error rate and in the case of 2-class problems, Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) are the most common performance metrics. ROC curves plot the sensitivity (also known as Type I error or True Positive Rate) which is the rate the event of interest is predicted correctly for all samples having the event against the specificity (where 1-specificity is the Type II error or False Positive Rate) which is the rate that nonevent samples are predicted as nonevents[11].

Variance refers to the flexibility of a model or how sensitive the model is to variations in the training set. Generally speaking, as model flexibility increases, the bias decreases (more accurate) and the variance increases (more sensitive).

[8] Everitt, B. (2010). Multivariate Modeling and Multivariate Analysis for the Behavioral Sciences. Boca Raton: CRC Press.

[9] Casella, G., Fienberg, S., & Olkin, I. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer Science and Business Media.

[10] Montgomery, D., Peck E., & Vining G. (2012). *Introduction to Linear Regression Analysis (5th Edition).* New Jersey: John Wiley & Sons Inc.

[11] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer Science and Business Media.

The effect of high variance is that models can become overfit. An overfit model is one where the training error is very low, but the test error is high. Test error follows a U-shape as flexibility increases. Depending on the application, a good model is one where a balance is found in the tradeoff between bias and variance.

Another trend with model flexibility is that the Explainability of how the model operates decreases as flexibility increases. For example, neural nets are very flexible approaches to modeling and can model complex nonlinear functions with low bias. They are, however, much more difficult to interpret and explain than a random forest regression model for instance. This paper focuses on CLS models which are more flexible approaches that tend to have low bias and high variance. One reason for this is that traditional parametric techniques are based on a set of assumptions about the underlying distribution of the input variable coefficients. These assumptions provide structure for Explainability. Clustering and decision trees can also have high Explainability as they output nice visuals which follow simple logic rules. Neural nets, however, assign a set of weights to input variables and iterate to convergence using a process called gradient descent. The difficulty in explaining these operations led to the common labeling of neural nets as "black boxes." Recent advances in image-recognition neural nets have shown promising results in highlighting areas that contributed the most to the classification outcome and reporting the associated level of confidence for the diagnosis[12].

A recommended practical modeling approach when a CLS model is desired for a supervised analysis is to create a series of models using different techniques which can range from linear regression to neural nets. The performance can be compared across the techniques, and based on the application and stakeholder feedback, a decision made on how to balance Explainability and performance.

Performance metrics are the key way to communicate and explain performance of a CLS model. There must be caution in interpreting and using the results of a well-performing model, however. The scope of low-bias neural nets is usually narrow and care should be taken when applying in the real world to ensure the use of the model is consistent with how it was trained. Also, models are typically "locked down" after validation so they are not continuing to learn off production data. As an example, recent neural net studies looking at retinal images that have narrow scopes such as triaging referrals from a training set of images with known retinopathy have better performance metrics than neural nets trained on images with no known retinopathy[5].

The Explainability of CLS performance starts with a discussion of performance metrics and comparison to targets based on risk-benefit analysis or comparison to known human performance levels. If possible, less flexible modeling methods can be used if performance is comparable and Explainability is higher. When using CLS models, care needs to be taken to ensure that the real-world application is reflective of the samples used for training the models.

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

5. Continuous Learning

Continuous Learning

How do updates impact the Explainability of the AI?

How can visibility into new data and updates help enable Explainability?

Just as humans gain experience over time, CLS systems are expected to be updated on a basis determined by the users: monthly, daily or even potentially in real-time. Explainability of the updates is crucial in keeping the user, regulator, and stakeholder informed of the evolution of the system.

There are two types of updates one can expect for an AI application: retraining or continuous training with a dataset of increased size; AI model or design change. The former is expected to be more frequent than the latter, but both should be considered in the context of Explainability of AI.

During the development of a product intended to analyze images for the presence of an abnormality, the manufacturer trains the algorithm on a diverse set of images, so it can handle anatomical variations and various disease severity and presentation. Once introduced onto the market the access to images will increase therefore allowing the algorithm to continue to train on additional anatomical variations or disease presentations not previously encountered in the training dataset. One can expect the performance of the algorithm to become more reliable in some circumstances or sub-population not otherwise well represented in the initial training dataset. Update to the training dataset will to lead to changes to the AI algorithm in terms of performance, potentially introducing bias, which needs to be communicated and explained to the user.

Data directly contributes to the performance of the algorithm and should be carefully managed and controlled. The quality and diversity of the data used to train or retrain an AI application should be subjected to proper data management, control and governance. Transparency on the nature of the training data and its evolution is necessary in order to explain the evolution of performance to the user. When the training dataset changes, the user should be provided with a way to understand how the system has evolved following retraining on a larger/different dataset or based on its continuous learning process.

Performance metrics are designed to explain how well the AI performs, re-training of the AI and associated improved performance should be explained to the user by describing in what way (e.g., updated labeling, disclosure statement, etc.) the new data differs from the data previously used to train the algorithm and how the performance metrics have been impacted by the update. When new data is used one should consider whether the performance metrics previously identified will suffice to explain the performance variations. It is possible that the performance improvement is limited to a small sub-population for which performance were not previously provided and the metrics previously identified to characterize the system do not make the change apparent to the user.

At times the AI model will be updated. Explainability of the design change is related to design Explainability. The same principles of design Explainability should be applied to a design change, focusing on the new element of the AI model

and its impact on performance. The impact of a design change has to be explained to the user and at times to the regulator and other stakeholders as well.

Indeed, when a manufacturer makes a change to a medical device, the change must be assessed to determine if a new regulatory submission is needed. The FDA has published two guidance documents ("Deciding When to Submit a 510(k) for a Change to an Existing Device" and "Deciding When to Submit a 510(k) for a Software Change to an Existing Device") to help manufacturers assess whether a change is likely or not to trigger a new regulatory submission. These papers employ a risk-based approach considering whether the change modifies the risk associated with the device.

AI retraining and AI model update are likely to impact the risk of the device differently. Retraining on a larger and more diverse dataset is likely, if done adequately, to improve performance of the algorithm. While the manufacturer should have in place a mechanism to confirm that the performance increases, it is likely that the risk associated to the device have not changed during retraining which reduces the need for a new regulatory submission. An AI model change is a more profound change in the device and is more likely to be intended to expand the indications for use of the AI application. Expansion of indications for use usually triggers a need for a new regulatory submission.

It should be noted that off-label use may result in data outside the predefined curation or training criteria being included in a CLS update. This could lead to an adulterated product. If fact, if this retraining occurred without the manufacturer's control, the application could be considered 'remanufactured.'

Additionally, one should consider how to manage continuous learning differences across the globe. For example, if one facility has five devices and a large and diverse population, their devices will learn differently than a device in a facility across the globe where they have a small and consistent population.

6. Explainability and Risk

Scaling Explainability or Trustworthiness to Risk

One aspect in determining how much Explainability or Trustworthiness is needed is to look at the risks involved if the application is wrong in its conclusions / recommendations.

In 2014, the International Medical Device Regulators Forum (IMDRF) published "Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations", which offered a categorization scheme based on the criticality of the patient and the significance of the role that the SaMD application has in care for that patient. The following categorization table is from that paper:

State of Healthcare situation or condition	Significance of information provided by SaMD to healthcare decision		
	Treat or diagnose	Drive clinical management	Inform clinical management
Critical	IV	III	II
Serious	III	II	I
Non-serious	II	I	I

Where the significance columns are defined as:

Treat or diagnose	Treating and diagnosing infers that the information provided by the SaMD will be used to take an immediate or near-term action: • To treat/prevent or mitigate by connecting to other medical devices, medicinal products, general purpose actuators or other means of providing therapy to a human body • To diagnose/screen/detect a disease or condition (i.e., using sensors, data, or other information from other hardware or software devices, pertaining to a disease or condition).
Drive clinical management	Driving clinical management infers that the information provided by the SaMD will be used to aid in treatment, aid in diagnoses, to triage or identify early signs of a disease or condition will be used to guide next diagnostics or next treatment interventions: • To aid in treatment by providing enhanced support to safe and effective use of medicinal products or a medical device. • To aid in diagnosis by analyzing relevant information to help predict risk of a disease or condition or as an aid to making a definitive diagnosis. • To triage or identify early signs of a disease or conditions.
Inform clinical management	Informing clinical management infers that the information provided by the SaMD will not trigger an immediate or near-term action:

	• To inform of options for treating, diagnosing, preventing, or mitigating a disease or condition. • To provide clinical information by aggregating relevant information (e.g., disease, condition, drugs, medical devices, population, etc.)

There are a variety of potential applications for AI in healthcare. Some applications will be assistive in nature, some will be augmentative, and eventually there will be fully autonomous systems. It is clear that the more significant the role is to the decision-making process, the higher the risk. Other industries have also looked at risk stratification using various levels of autonomy – a fully autonomous system typically has a higher level of concern than one in which a person makes a final judgment.

In the book *Army of None: Autonomous Weapons and the Future of War*, the author offers three levels of human participation:

1. The machine performs a task and then waits for the human user to take an action before continuing
2. The machine can sense, decide, and act on its own. The human user supervises its operation and can intervene, if desired
3. The machine can sense, decide, and act on its own. The human cannot intervene in a timely fashion.

This raises the question if the IMDRF "Treat or diagnose" should be further subdivided into additional categories to clarify the role of the human in the loop.

State of Healthcare situation or condition	Significance of information provided by SaMD to healthcare decision				
	Treat or diagnose w/no intervention possible	Treat or diagnose w/Override	Treat or diagnose w/Approval	Drive clinical management	Inform clinical Management
Critical	TBD	TBD	IV	III	II
Serious	TBD	IV	III	II	I
Non-serious	IV	III	II	I	I

These proposed additional categories shown begin to help categorize the numerous types of use-cases that AI enables. In the final analysis, this is likely only the starting point of a framework that allows all stakeholders to discern the extent of Explainability needed for a particular solution or use case. Subsequent development of this framework will undoubtedly be forthcoming.

6. Conclusion

This paper was developed as a follow up to the "Perspectives for Good Practices in Continuously Learning AI Systems in Healthcare" published in August 2018 during the Xavier Health AI conference. As the title indicates, our focus here was to establish a broad discussion around Explainability and Trust in AI applications.

The notions of Explainability, Trust, and Trustworthiness have taken on special meaning in the developing vocabulary for AI applications in the last year. While some of these terms are used interchangeably by many, standard developing organizations such as the ISO International and others have made great strides in clarifying these terms for the industry at large. For our purposes, Explainability is quite simply the extent to which the workings of a systems or its output can be sufficiently explained to stakeholders.

As this paper describes, there are many approaches to providing satisfactory levels of "explanation" ranging from the type of AI technology being utilized, to the process and method being used for providing this level of explanation. As a specific example, DARPA's XAI initiative has provided a great foundation for describing the characteristics of an "explainable system" vs. a more traditional neural network-based system, which is undoubtedly going to benefit a wide range of stakeholders. Correspondingly, the Explainability requirements within EU's General Data Protection Regulation (GDPR) have undoubtedly catalyzed much of the recent dialog and evolution of this topic.

This paper also delves into required levels of Explainability based on the associated risk of AI-based application. In other words, how should the extent of Explainability change based on whether the solution Treats or Diagnoses vs. Drive Clinical Management vs. Inform Clinical Treatment. This becomes increasingly important for those applications where the extent of Explainability negatively impacts the performance of the system.

In the final analysis, Explainability remains a rapidly evolving topic. Our notion of how we define or address Explainability is likely to get significantly refined over the next three to five years. We hope that our approach to drive this discussion across proven and well-understood systems constructs will help accelerate our pathway forward.

Appendix I – Glossary

It should be noted that in the AI/CLS/ML field, terms are often used casually and interchangeably. For example, "adaptive learning" and "continuously learning" may be used to describe the same concept. For purposes of clarify, the following definitions are used in this paper.

Terminology

The following are identifiers that will be used in this paper that refer to the broader category and inclusion of related technologies and stakeholders:

ADVANCED BROAD-BASED ANALYTICS (ABBA)
Analytics based on a large volume of data as well as a variety of different types of data.

ALGORITHMS (CLUSTERING, CLASSIFICATION, REGRESSION, AND RECOMMENDATION)
A set of rules or instructions given to an AI, neural network, or other machine to help it learn on its own.

ARTIFICIAL INTELLIGENCE (AI)
A machine's ability to make decisions and perform tasks that simulate human intelligence and behavior. Alternatively – 1. A branch of computer science dealing with the simulation of intelligent behavior in computers. 2. The capability of a machine to imitate intelligent human behavior (source: Merriam-Webster)

NEURAL NETWORK
A learning model created to mimic some aspects of the human brain to solve tasks that are too difficult for traditional computer systems to solve.

AUGMENTED INTELLIGENCE, also known as INTELLIGENCE AUGMENTATION (IA)
Systems that are design to enhance human capabilities. This is contrasted with Artificial Intelligence, which is intended to replicate or replace human intelligence.

CLASSIFICATION
The problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

CLUSTERING
Clustering algorithms let machines group data points or items into groups with similar characteristics.

CONTINUOUSLY LEARNING SYSTEMS (CLS)
Continuous Learning Systems are systems that are inherently capable of learning from the real-world data and are able to update themselves automatically over time while in public use.

DECISION TREE
A tree and branch-based model used to map decisions and their possible consequences, similar to a flow chart.

DEEP LEARNING
The ability for machines to autonomously mimic human thought patterns through artificial neural networks composed of cascading layers of information.

EXPLAINABILITY
A human-comprehensible explanation of how a 'black box' model is statistically likely to have come to a specific recommendation. Source: *Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care,* Duke Margolis Center for Health Policy,

INTERPRETABILITY
A human-comprehensible explanation of exactly how the model combines and uses inputted data to come to a specific recommendation. Also referred to by computer scientists as "model transparency". Source: *Current State and Near-Term Priorities for AI-Enabled Diagnostic Support Software in Health Care,* Duke Margolis Center for Health Policy,

MACHINE LEARNING
A facet of AI that focuses on algorithms, allowing machines to learn and change without being programmed when exposed to new data.

NATURAL LANGUAGE PROCESSING
The ability for a program to recognize human communication as it is meant to be understood.

TRUST
firm belief in the reliability, truth, ability, or strength of someone or something

TRUSTWORTHINESS
Quality of being dependable and reliable. Source: ISO 17068:2017

EXPLAINABILITY
Information systems that collect and analyze rea

INTERPRETABILITY
Information systems that collect and analyze rea

SUPERVISED MACHINE LEARNING
A type of Machine Learning in which training datasets contains the desired or targeted outputs so that the machine can be trained to generate the desired algorithms similar to the way a teacher supervises a student.

UNSUPERVISED MACHINE LEARNING
A type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses (e.g. cluster analysis.)

Appendix II – Project Background

During the first Xavier Artificial Intelligence Summit in August 2017, a working team of FDA officials and industry representatives was formed to identify successful practices for evaluating systems that continuously learn. A whitepaper on Good Machine Learning Practices and a companion checklist was created by that team and published in August 2018. At the 2018 Summit, another working team was formed to develop this paper regarding Explainability and trustworthiness.

Appendix III – Related Initiatives – FDA PEAC Project

The Patient Engagement Advisory Committee to the Food and Drug Administration (FDA) has been formed to make recommendations on the topic "Connected and Empowered Patients: e-Platforms Potentially Expanding the Definition of Scientific Evidence." The recommendations will address how FDA can leverage patient-generated health data, such as social media, sensors and patient-driven registries, to better engage patients and consumers as empowered partners in the work of protecting public health and promoting responsible innovation. Social media and other web platform enablers are facilitating the growth of virtual patient communities. Increasingly, patients and healthcare consumers are using these platforms to share their health experiences and seek information from other patients and consumers, rather than their healthcare providers alone. This overall goal of engaging patients works towards their acceptance and adoption of Digital Health and AI based apps and solutions in healthcare.

The figure below expands the use of secondary, shared and published data. Expansion of secondary data sources is necessary to reduce time to adoption, allowing parallel processes to occur in real time.

Sources of Secondary Data For Imputs	Examples
Electronic Health Records	- Pharmacy, Laboratory & Clinical Databases, and Unstructured Data
Hospital Billing Data	
Payer Claims Data	
Health Maintenance Organizations	- Pharmacy, Laboratory and Clinical Databases, Paper-Based & Unstructured Data
Published Literature	- Meta-Analyses, Patient Reported Outcomes, NIH Funded Data
Patient Registries	
National Survey Databases	- Medical Expenditure Panel Survey, formerly National Medical Expenditure Survey
Sensor Data	- Consumer wearable fitness devices, locaters and cameras - Implantable and external medical devices
Social Media	- Structured and Unstructured Data

Appendix IV – Additional Considerations for Explainability in Presentation and Adoption

Presentation Considerations

The following represents some highlighted Considerations and Questions for developers and business managers in "presenting" the AI application.

Considerations

- What types of information do we present to the user (e.g. patient or clinician) in a real-time setting? Should there be a way to delve deeper in real time or look at accuracy information?
- Considerations for the cognitive burden of the user -- how can we leverage standardization to reduce the burden?
- Need to avoid information overload – e.g., Airbnb, Facebook. Represent the data elements in ways that people would understand. How do we make it easy to understand for end users?
- Understand the user's current workflow and how to seamlessly fit into (or replace) it
- Ensure language communicated to patients about how the data being collected will be used is clear and understandable
- Explain how data is collected, used and applied in the various situations and workflows in Healthcare. What is presented to Radiologists can be very different from that presented to a family practitioner.
- Ensure data collected is de-identified
- Ensure a high level of integrity and security for the collected data

Questions

- What are the different methods for delivering an output that incorporates Explainability?
- For user-facing applications, what approaches might be considered to deliver confidence and transparency in output?
- How confident are users in data sources? Provide source information that is credible.

Adoption Considerations

The following represents some Considerations and Questions for developers and business managers in "the adoption of" the AI application:

Considerations

- How to design the application to fit within user's current workflow, or replace the workflow? How will the application improve patient care without increasing workload?
- Additionally, if the application slows down or interrupts the user's workflow, they are likely to resist adoptions.
- If the application does not fit within the user's mental model, they are likely to make mistakes when using it (i.e. use errors).
- If the information presented is too complicated or too voluminous for the user to digest, then the risk of it not being trusted or adopted will increase and the potential for the actual use of the information will be decreased.
- Include discussion of RWE and impact to patient outcomes. Publication, education, patient centered research. Include de-mystifying the technology and address the fear factor.
- End users current workflow and mental model of the system
- Regulatory impacts of CLS (what is enough change to require 510(k)). When the algorithm changes, does the manufacturer need to re-verify/re-validate? What level of change would trigger this?
- Build more trust between patients, industry, vendors and the government
- Consider socioeconomic disparities in access and use of certain technologies such as sensors and smartphones
- Encourage the involvement of patient organizations in the collection of data
- Ensure that end users (whether clinicians, nurses, or patients) are educated on the use of their devices, how the data is being collected, and that the language is clear
- Ensure that the patients are true patients and not reflective of influencers or other "fake patients"
- Develop standard data elements and formats in which data is collected and reported
- Develop integration requirements for product application and other systems
- Develop entertaining educational communications by leveraging social media (Vines, YouTube, Facebook, etc.)
- Consider international differences in adoption and presentation

Questions

- Practices and processes that are in place to protect patients' privacy. Including, language that informs patients about some possible uses of their data in medical research, clearly explains informed consent, is transparent, and informs patients how to access their data.
- How to inform individuals about the use and application of their data.
- Patients should have access to their data to use and share.

- What level of explanation is necessary to drive confidence in the output of the system to different types of stakeholders?
- What level of training is necessary for the end users? Should this be documented/mandated by regulatory bodies? The goal is to reduce user errors to zero.
- What are the different methods for delivering an output that incorporates Explainability?
- For user-facing applications, what approaches might be considered to deliver confidence and transparency in output? E.g., Human Factors?
- How confident are users in data sources? What are the rules for curation of learning, test, QC, and real-time patient data?
- What is good enough?
- Who owns the data? In many states, patients do not own their medical records[13]

[13] https://www.washingtonpost.com/national/health-science/who-owns-your-medical-data-most-likely-not-you/2018/11/23/28785efc-e77d-11e8-a939-9469f1166f9d_story.html?utm_term=.10c4feda0f5e

Appendix V – Trust in AI Solutions

To increase the speed of innovation and adoption, it is necessary to create parallel processes, by specialized teams with expertise in the subject matter and disease of interest. Three continuous teams are Human Factors Design team, the Research Publication and Presentation team, and the Business Development team. Four parallel audiences are targeted for adoption, each with unique interests: Reimbursement, and Formulary decision-makers, Early Clinical users, Mid-range Clinical Users, and Late Adopters.

Current Environment for Clinicians and Patients

Presentation & Adoption acceptance starts with the physician and ethical considerations and trust to use the AI tools in their practice and in the treatment of patients. It ends with the patient. There is a lot of hype around AI technology and the potential it holds. Applications currently in use have not adopted the more sophisticated forms of Continuous Learning Systems. This crawl/walk/run approach is the natural evolution of any technology and how it is adopted by both users and society. In some respects, we are at the Walk stage in the maturity of the technology given that AI us ubiquitous in our daily lives and the way that AI surrounds us and, in some cases, we don't even know it. Alexa, Pandora, Internet shopping and advertising as well as the financial institutions all use AI to improve our experience and convenience as they learn from our behavior and adapt to our interests and preferences. New vehicles are ever increasingly using sensors and AI algorithms to improve safety and autonomous driving. Likewise, in current healthcare applications AI becoming more prevalent in the area of medical imaging across a broad sector of diagnosis and treatment of diseases.

Focus on improved efficiency, for instance, and how these evolving technologies can help make that a reality particularly by fitting seamlessly into the already busy and complicated workflow of health care providers. C-Suite executives care about improving efficiency because it can save time and, ultimately, save money. If you want to eventually integrate AI into your standard workflow, speaking about it in those terms is the way to ensure you get the needed investments[14].

Seamlessly and efficiently integrating AI can also reduce the number of use errors and thus reduce any consequential patient harm.

In order to accelerate the adoption of AI in medicine, clinicians will need to embrace the technology. Therefore, the way that the technology is presented to the clinicians will have to engender trust.

Clinicians rely on mechanistic explanations that correlate observations (inputs) with results (outputs). They use patient history and physical exams, laboratory analysis, radiological imaging and evidence based critical thinking to deduce the cause of a condition and make a clinical decision. When presenting AI to clinicians, one has to keep the pathway in which a physician makes decisions in mind.[15]

[14] https://www.radiologybusiness.com/sponsored/9667/topics/imaging-informatics/rsna-2018-hype-ai-radiology-paul-chang
[15] Shaywitz, David, *AI Doesn't Ask Why—But Clinicians and Drug Developers Want to Know*, Forbes (November 2018). https://www.forbes.com/sites/davidshaywitz/2018/11/09/ai-doesnt-ask-why-but-clinicians-and-drug-developers-want-to

Clinicians serve as the gateway for AI to patients. Thus, clinicians need to be confident that the AI is safe and performs as well as or better than the current evidence-based treatments. Also, physician must have enough depth of understanding of the AI to be able to explain it to medically savvy patients. The comprehension has to be at the level that the physician will feel comfortable obtaining informed consent from patients for the use of AI.

The Matter of TRUST!

Trust, in terms of humans and automation, can be defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." This is extremely important in the health care environment where there is a lot of uncertainty, increased by new complex technology, and a lot of vulnerability, when asking that technology to play a crucial role in patient-related outcomes.

Establishing the right amount of user trust is very important in automation to avoid complacency (over trust) or the "cry wolf effect" (under trust). The amount of feedback and type of feedback are important when considering under-trust and over-trust in AI. To mitigate under-trust or mistrust, where the user does not believe the information of the system and thus does not act on it, there would need to be transparency to the user of the accuracy of the system. Likelihood displays, which show how accurate the system's alarms are, help the user understand how much to trust the system. Additionally, if the system shows the user how it is making its choices (what algorithm it is using, what factors it is taking into account, if continuous learning is occurring) this will help the user decide how much to trust the system.

Equally dangerous is complacency, or over-trust, where the user is over-dependent on the system and does not question the accuracy of the system. Not only does this lead to poor choices if the system performs inaccurately, but also if the system malfunctions or stops working, the user is likely to have lower situational awareness and not know how to pick up clinical action.

Additional information about this topic can be found in the book *Engineering Psychology and Human Performance*, fourth edition, Wickens et al. and in the paper "Trust in Automation: Designing for Appropriate Reliance", https://user.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf

Quality of Data

Post market regulatory activities including advising manufacturers on the development of post market studies (such as defining the study design, patient population, and outcomes of interest), performing surveillance for adverse events, issuing recalls, and communicating signals to the public.

Patient-generated health data should be used, to inform these post market processes for medical devices by collecting data on surveillance for adverse outcomes, issuing recalls and signal management.

Patients should be informed of post market performance of the AI applications that are being used and/or have been used in their diagnosis and treatment in a timely manner.

Data Quality, and Data Governance in general, does not seem to be a high priority topic among medical device regulators. While clinical data validity is of course critical for providing evidence for any change in clinical practice, overall data governance best practices for software development and machine learning do not seem to be held to the same level of scrutiny. It took the 2008 financial crisis for banking regulators to demand fundamental changes in data governance practices that help ensure that banks and insurers understand what their data means, where it comes from,

what quality checks are in place, and how it is all being calculated. While the Sarbanes-Oxley Act had previously attempted to provide guidance, it was inadequate as shown by the subsequent financial crisis. The Basel committee's BCBS 239 regulation is a far more robust and mature regulatory framework for this discipline. GDPR is another example of a robust data governance regulatory framework that has been implemented after egregious practices. For AI in the medical community, any preemptive efforts to ensure proper data governance across organizations would be beneficial.

Models

Human Factors pre-market studies can give insight into how these systems will be used and what the likely mistakes may be. Building confidence in the classifiers is critical at the early stages. It could be argued that explaining too much to too many people can decrease their confidence in the device. Ideally, it should "just work" and, as explained above, provide diagnoses as the provider is accustomed to receiving. We can build confidence by ensuring proper data inputs; for example, demonstrating that high quality images were used in the training set guiding the user to capture high quality images, and explaining how to achieve that, will help ensure that the user gets more skilled at providing the necessary data inputs.