Evaluations of AI Applications in Healthcare Study Guide

Study Guide

CONTENTS

MODULE 1: AI IN HEALTHCARE

LEARNING OBJECTIVES

- Recognize why we need AI in healthcare what understand what are the right questions we can ask AI to help solve
- Describe the breadth of AI in healthcare, from the molecular level to the population level.
- Recognize the different categories of AI in medicine (biomedical research, translational research and medical practice)
- Explain why AI evaluations need to move beyond model accuracy

AI IN HEALTHCARE

Artificial Intelligence (AI): Involves the development of computer algorithms to perform tasks typically associated with human intelligence

AI spectrum of learning:

- Machine learning
- Representation learning
- Deep learning
- Natural language processing

Algorithms: A mathematical technique, generally developed by statisticians and mathematicians for a particular task, such as unsupervised learning or reinforcement learning

Models: Well-defined computations formed as a result of an algorithm that takes some value, or set of values, as input and produces some value, or set of values as output

AI solution: Refer to an evaluated and validated model

AI has the potential to provide high performance data-driven medicine, optimize care trajectories, suggest the right therapy for the right patient and improve the process of clinical assertions and decision making.

The best way to understand AIs potential use in healthcare is to think about its applications in three separate categories:

- Biomedical research: AI is assisting in automated experiments, automated data collection, gene function annotation, literature mining
- Translational research: AI is assisting in areas such as biomarker discovery, drug-target prioritization, and genetic variant annotation
- Medical practice: AI used for disease diagnosis, treatment selection, patient monitoring, and risk stratification models.

Reasons why AI is needed in medicine:

- **AI and Data Synthesis-** AI can make sense of the overwhelming amount of data associated with a single disease or a single patient to highlight the relevant information needed to best guide the treatment and care for each individual patient.
- AI can improve **clinical reliability** and be used to help identify relevant information
- AI tools can tackle tedious, mundane tasks because they don't suffer from fatigue, distractions, or moods like their human counterparts. Therefore, AI tools can be useful for **reducing errors related to human fatigue**.
- **Patient-clinician engagement is another important area where we see AI solutions emerging.**
- AI can **improve patient outcomes** by prioritizing patients in more urgent need and by recommending individualized treatments that account for a patient's unique characteristics, which often fall outside of the selective clinical trial evidence.
- AI has the potential to improve resource utilization and efficiency, thereby **reducing overall healthcare costs**.
- Ai is being used to **improve diagnostics.** AI can quickly and more accurately spot signs of disease in medical images such as MRIs, CT scans, ultrasounds and x-rays.

- AI can **help physicians understand patients' values and goals.** AI can assist by analyzing structured and unstructured patient data and present insights for physicians' consideration and to aid in shared decision making.

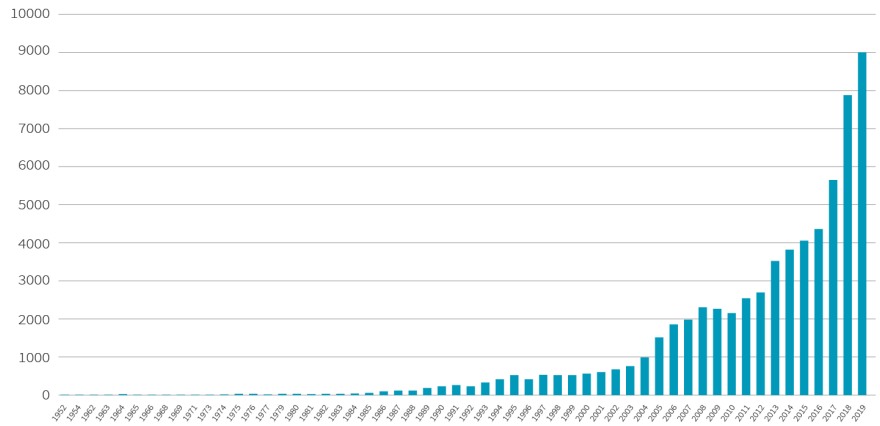
We can think about AI solutions in medicine across the domains of biomedical research, translational research and medical practice.

A non-exhaustive list of some recent AI applications in medicine that demonstrate how AI algorithms can benefit patients, doctors, and researchers through making diagnosis, treatment, discovery, and the practice of medicine faster, more accurate, and more efficient:

- Image analysis - AI is being used in clinical practice for disease diagnosis, AI solutions can reduce errors related to human fatigue and Improve diagnostics.
- Drug discovery - AI development under the translational research branch. Key classification features are gene essentiality, mRNA expression, DNA copy number, mutation occurrence and protein–protein interaction network topology.
- Patient risk stratification (or Population level segmentation) - AI is used in clinical practice for risk stratification.
- Risk of 30-day readmissions – Another example of AI used in clinical practice for risk stratification.
- Basic biomedical research - AI is being used in basic research for automated experimentation
- Home videos for autism diagnostics

These examples provide a glimpse of the broad range of AI applications we see emerging in medicine and the opportunities to improve diagnostics, care delivery, access to care, and patient outcomes.

GROWTH OF AI IN HEALTHCARE



Both literature and media show that there is a hype around AI and this is coupled with an explosion of AI applications in healthcare. Recently, there has been a drastic increasing trend in medical literature related to "artificial intelligence" and "healthcare" or "medicine".

Most research:

- Provides an overall description of the clinical problem
- Includes the architecture of the proposed model
- Provides the accuracy or precision of the model
- Provides thoughts about how their AI solution COULD be used in the healthcare setting

Think beyond simple accuracy or precision as a measure for AI models because these measures do not tell you anything about the impact of your model in the healthcare setting.

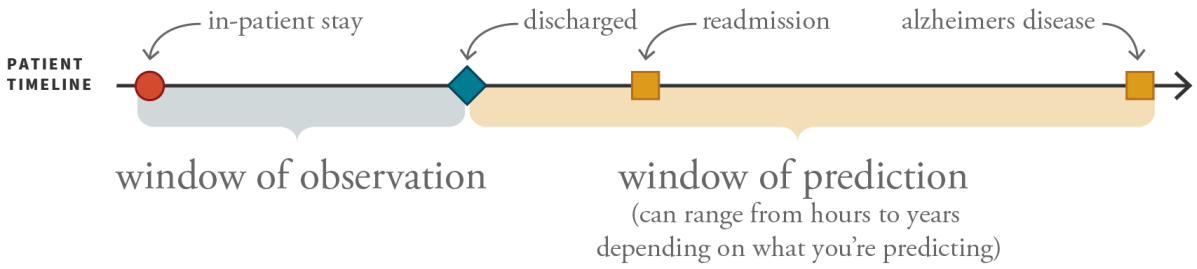
Often researchers focus on a model without considering if the model output can be mitigated or will have an impact on clinical care.

A systematic review identified articles predicting hospital readmission risk:

- Identified about 8,000 citations
- Included 30 studies that met the inclusion criteria
- Most models relied on retrospective administrative data
- Had poor discriminative ability (with a c-statistic less than 0.8)
- Only 1 study specifically addressed preventable readmissions

The ability to predict who is at high risk for hospital readmission may not be clinically useful if the model does not convey anything that the team does not already know. However, if the model contained data on discriminatory actions during the inpatient stay that were associated with low risk of readmission then this information could be taken into account to improve the utility of the AI solution. However, most algorithms are developed without considering the clinical utility, feasibility, or the overall impact of implementing the AI solution.

HOW TO KNOW IF AN AI MODEL IS GOOD



An example of a longitudinal data for a patient – There will be a window of observation, where you are collecting and aggregating data. This observation window could be the inpatient hospital stay. Then you set a time where you will start predicting an event. In our example, this could be time of hospital discharge. You make a prediction and that event can happen within the next hour or in the next N years, depending on what you are predicting. If you are predicting patient deterioration, it may happen within minutes or hours after the observation period. However, if you are predicting Alzheimer's Disease based on genotyping and family history, your event might not occur for years, maybe even decades. So, this gives you your action window. And depending on the prediction, this action window can be categorized as acute or long-term.

One must think about whether the action to my output has to happen immediately, or is it something that's going to happen over the next year, five years, or 10 years? Studies have shown that, regardless of acute or long-term actions, early warning lead times give more opportunities for action.

Interestingly, almost no effort goes into the action side of the equation. You make a prediction; you publish the paper and then you are done.

What we need to think about is how do we move beyond predictions and start thinking about the action side of the evaluation: How will the prediction be used and by whom - so that we can build better models and make sure they get integrated into clinical care.

Start thinking beyond predictions and to do this, we will start with a framework that allows us to conceptualize **outcome-action pairing (OAP)**. In OAP, the outcome is the purpose of the AI model: disease diagnosis, risk stratification, or event prediction. The action is the step that can be taken based on the outcome that will improve medical care. A different treatment pathway based on risk, a new treatment based on a diagnosis, or scheduled follow-up can with a primary care physician to prevent a readmission.

To evaluate an AI solution beyond its predictive value, we need to think about a framework that can be systematically applied across the broad range of AI solutions emerging in healthcare. We need a framework that provides criteria to evaluate the utility, feasibility, and overall clinical impact of an AI solution.

- **Utility**: Refers to the purpose of the model and it is what matters to the patient most
- **Feasibility**: Refers to what is needed to implement the AI solution in the healthcare setting
- **Clinical Impact**: The overall effect that the AI solution can have on clinical care, patient outcomes, and care standards

Artificial Intelligence spans the breadth of healthcare and there are so many opportunities to use AI to improve upon or augment healthcare delivery. AI can identify patterns in the expanding, heterogeneous data sets in healthcare to create models that accurately classify, predict or recommend actions. However, realizing the potential benefit of AI solutions for patients in the form of better care requires rethinking how we evaluate AI solutions A framework for rigorously evaluating the performance of a model in the context of the subsequent actions it triggers is necessary to identify AI solutions that are clinically useful.

CITATIONS AND ADDITIONAL READINGS

Coquet, J., S. Bozkurt, K. M. Kan, M. K. Ferrari, D. W. Blayney, J. D. Brooks, and T. Hernandez-Boussard. 2019. "Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients." *J Biomed Inform* 94: 103184.

Hao, S., Y. Wang, B. Jin, A. Y. Shin, C. Zhu, M. Huang, L. Zheng, J. Luo, Z. Hu, C. Fu, D. Dai, D. S. Culver, S. T. Alfreds, T. Rogow, F. Stearns, K. G. Sylvester, E. Widen, and X. B. Ling. 2015. "Development, Validation and Deployment of a Real Time 30 Day Hospital Readmission Risk Assessment Tool in the Maine Healthcare Information Exchange." *PLoS One* 10(10): e0140271.

Jeon, J., S. Nim, J. Teyra, A. Datti, J. L. Wrana, S. S. Sidhu, J. Moffat, and P. M. Kim. 2014. "A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening." *Genome Med* 6(7): 57.

Kansagara, D., H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani. 2011. "Risk Prediction Models for Hospital Readmission: A Systematic Review."

Matheny, M. E., D. Whicher, and S. Thadaney Israni. 2019. "Artificial Intelligence in Health Care: A Report From the National Academy of Medicine." *JAMA*.

Noack, M. M., K. G. Yager, M. Fukuto, G. S. Doerk, R. Li, and J. A. Sethian. 2019. "A Kriging-Based Approach to Autonomous Experimentation with Applications to X-Ray Scattering." *Sci Rep* 9(1): 11809.

Rajpurkar, P., J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren. 2018. "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists." *PLoS Med* 15(11): e1002686.

Tariq, Q., J. Daniels, J. N. Schwartz, P. Washington, H. Kalantarian, and D. P. Wall. 2018. "Mobile detection of autism through machine learning on home video: A development and prospective validation study." *PLoS Med* 15(11): e1002705.

Yu, K. H. and I. S. Kohane. 2019. "Framing the challenges of artificial intelligence in medicine." *BMJ Qual Saf* 28(3): 238-41.

MODULE 2: EVALUATIONS OF AI IN HEALTHCARE

LEARNING OBJECTIVES

1. Describing a framework for evaluating AI applications in healthcare
2. Understanding clinical utility and outcome-action pairing in AI solutions
3. Recognizing the many different aspects of an action to an AI solution.

A FRAMEWORK FOR EVALUATION

Two important aspects of evaluation: stakeholders and beneficiaries. These are important attributes to consider in the development, design, and deployment AI solutions.

Stakeholder Involvement:

- Important to understand what stakeholders should be involved throughout the design, development, evaluation, validation and deployment of an AI solution.
- Involves knowledge experts, decision makers, and end-users

Beneficiary:

- Involves understanding who the AI solution is made for or who it will be used by
- Could be a provider, patient, hospital, payer

Stakeholders and beneficiaries are components that need substantial thought and consideration in the development, design, and deployment AI solutions.

Clinical Utility: Relates to its applicability and impact on the healthcare system

- It requires identifying the beneficiary of the AI solution and understanding what action can be taken based on the model outcome that will improve for the beneficiary
- Two components (action and outcome) will allow you to better understand the problem addressed by the AI solution, and if it is a problem worth solving with AI

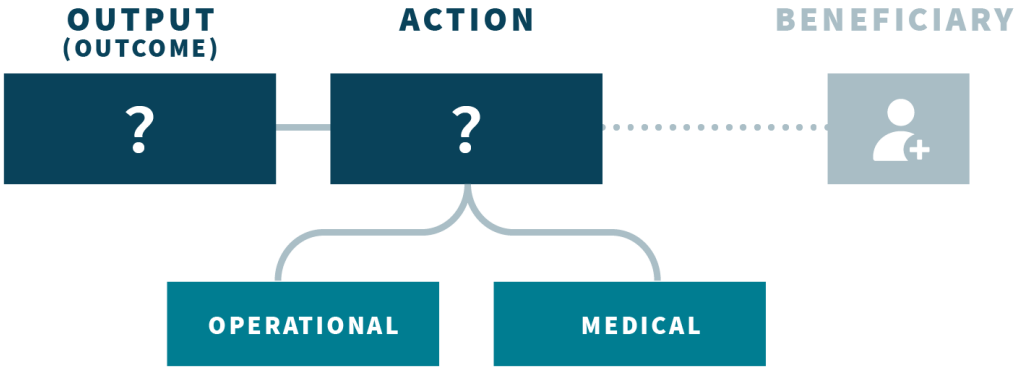
Considerations when evaluating an AI solution:

- What action can be taken based on the model output?
- Does a mitigating action (or therapy) exist?
- What is the ability to intervene? Who needs to take the action?
- What is the lead time offered by the prediction?
- What are the logistics and costs of the intervention?
- What are the incentives for acting on the output?

Start with the problem you are trying to solve with AI and then ask, is that problem worth solving. To help you answer this question for any AI solution, we will start by analyzing an action that can be paired with the model outcome - something we refer to as the "output - action pairing" or what the cool people will call the "OAP"

Utility Assessment: For every good AI solution (or prediction), there should be the ability to act upon or mitigate the output

OUTCOME: ACTION PAIRING



In OAP, the **outcome** (or output of the model) is the purpose of the AI solution, for example, a disease diagnosis, risk stratification, or event prediction. An **action** is a step that can be taken based on the outcome that will improve medical care. When we think about evaluating an AI solution, we must understand the outcome and know if there is a mitigating action that could change this outcome. This is the basis of the outcome-action pairing framework.

It is important to remember that for every good AI solution there is the ability to act upon or mitigate the output.

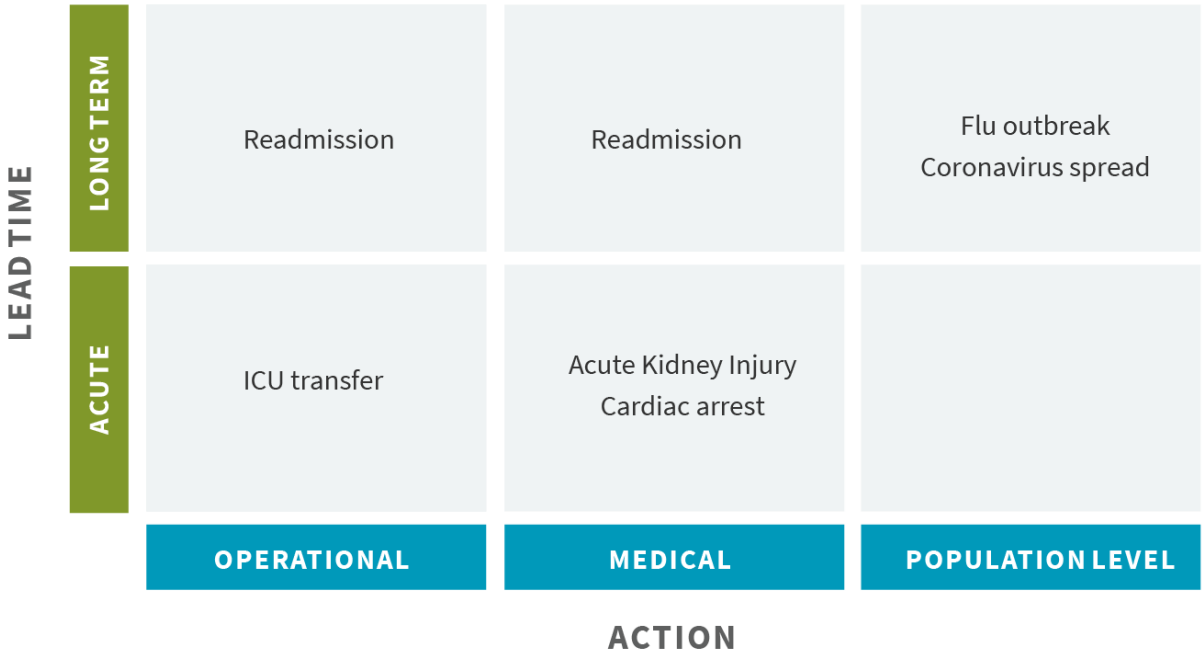
When evaluating an AI solution, it is also important to think about the lead time your action needs. One must think about whether the action to my output is acute (it has to happen immediately), or is it long-term?

Studies have shown that, regardless of acute or long-term actions, early warning lead times give more opportunities for action. So, the further in advance I can make my prediction the more time one would have to respond with the action to that output.

Lead-Time provided by the AI solution can directly impact its clinical utility in the healthcare system.

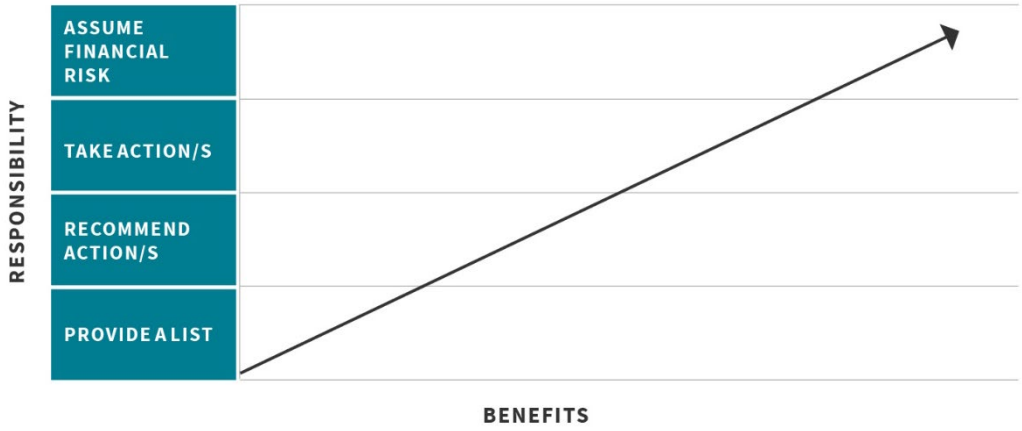
Once the action is decided based on AI outcome, the next question to ask is: What type of action should you take? You must clarify the type of action to evaluate the feasibility to implement the AI solution at point of care.

We need to understand if the action is **operational** or **medical** - and an AI solution can be developed to have either action.



This is a simple representation of the concept, with action types on the y-axis and action lead-times on the x-axis. What if my AI solution predicts ICU transfer? Where would this fall on the grid? This could be an acute operational action, as the hospital care team would need to act on this prediction by possibly 1) arranging the transfer; 2) finding a bed in ICU; and 3) identifying the ICU care team. If I am predicting cardiac arrest, this would be an acute medical action. Alternatively, if my AI solution predicts a hospital readmission, this could be either an operational or medical action. If the action is to schedule an appointment with a primary care provider, the action is operational and long-term.

A **population-level action** is related to AI solutions that predict public health events or other population outcomes, such as flu outbreaks or the coronavirus spread. If you are predicting flu outbreaks or vaccine efficiency, in addition to your local stakeholders, it is likely you would want regulatory agencies, and/or local, state and federal governments as stakeholders. It is important to keep in mind that depending on the type of action, different stakeholders will be needed.



For instance, your output from an AI solution can be the predicted risk of a 30-day hospital readmission and your action can be to provide a list of patients at risk. A minimal action is, you create a list of, for example 100 patients at highest risk for hospital readmission.

- Your Prediction: Risk of Hospital Readmission
- The Action: Provide a list of the 100 highest risk patients

You can send a list to somebody and hope for the best. For example, you can sell your list to the healthcare system and then it's up to the health system to do something with that information. And a lot of health systems and insurance companies will buy these lists and then provide that list to their disease management or care management teams. In industry, this is called a chase list.

If you go one level up on the action side, you might recommend an action. You need to understand what is going on and what the features mean in your model. A black box algorithm might not work with this action. Here is where interpretability and explainability usually are invoked.

The next step is that you figured out what action is needed to take and then you actually take the action - you execute and follow through.

Finally, if you're really sure and confident about the efficacy of your action plan and your ability to execute it, you might start assuming financial risk.

These are some examples of how we can think about outcome-action pairing and how this framework can be used to evaluate AI solutions. Outcome-Action pairing can be an effective way to evaluate how you or if you would deploy an AI solution.

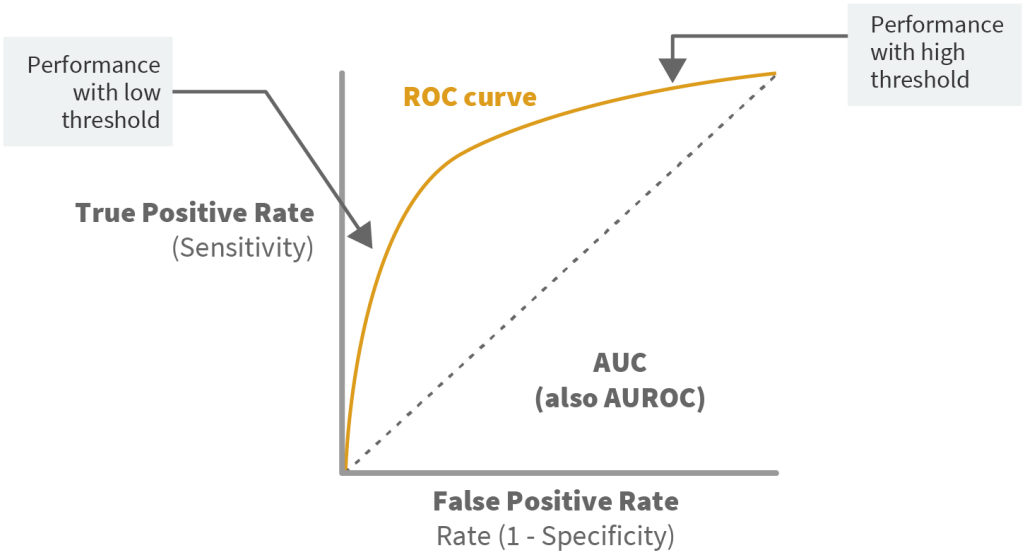
CLINICAL UTILITY

When we think about utility, we can also think about different measures that assess the number needed to benefit from the model. The value of the predictive model and its resulting actions can be conceptually divided into two components: number needed to screen and number needed to treat

- **Number needed to screen**: The number of people you need to screen to identify one "true positive"
- **Number needed to treat (NNT)**: The number of patients needed to be treated for one patient to benefit or the number of true positives you would need to take action on or treat for 1 patient to benefit
- **Number needed to harm (NNH)**: The number of people who received the intervention in question that would lead to just one person being harmed. With NNH, instead of looking at desirable outcomes, you are comparing the absolute risk increase of bad outcomes.

To evaluate clinical utility is to define and characterize the problem to be addressed by the AI solution and to determine whether that problem can be solved (or is worth solving) using AI

We often look at the **Receiver Operator Characteristics (ROC) curve** to evaluate an AI solution.



The ROC curve is a plot of the true positive rate against the false positive rate at different threshold settings. The true-positive rate is also known as sensitivity or recall. The false-positive rate is also known as probability of false alarm, which is 1 – specificity.

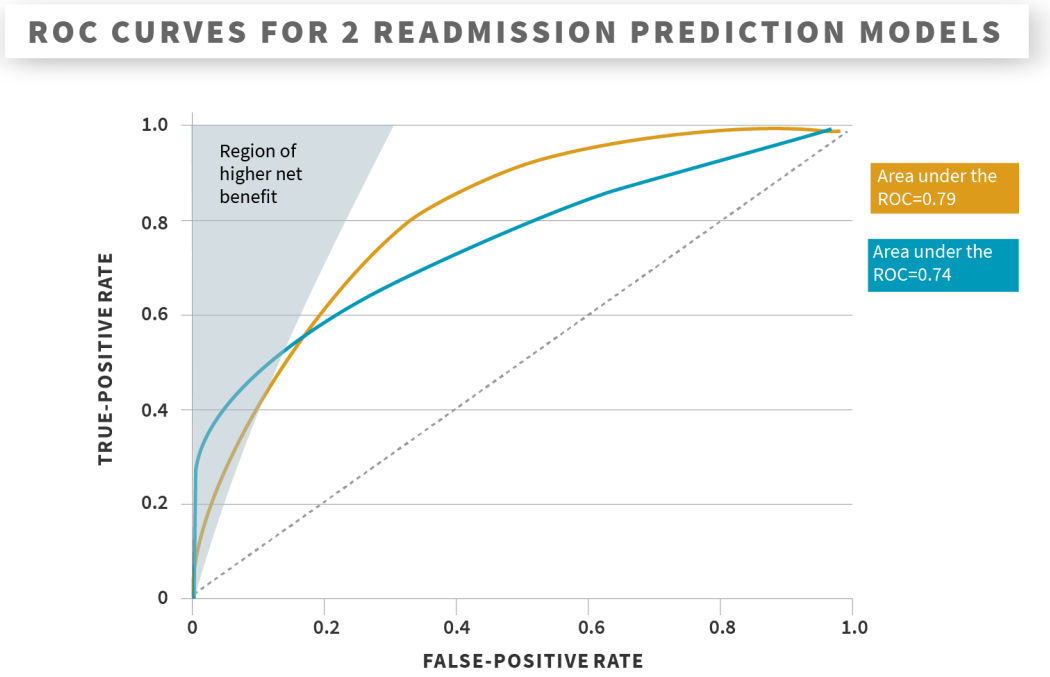
You need to get a cost and a utility for your true positives and you need the same information for your false positives. From this, you can estimate a zone on your ROC curve that if you set a

threshold, you would make better cost decisions than if you were to set a decision threshold - **a net benefit analysis**.

Generally, a **2-step process** of selecting a best model and then evaluating whether the model is helpful is used. This can be misleading because in machine learning hundreds of computerized models are created from which 1 is selected during the process of learning.

A decision curve analysis takes the threshold probability of an event where the relative costs of a false-positive and a false-negative prediction are taken into consideration. This theoretical relationship can be used to derive the net benefit of the model across different threshold probabilities - which generates a "decision curve."

A decision curve can be used to derive the net benefit of the model across different threshold probabilities.



Given estimated costs and benefits of the actions possible to mitigate a readmission, there is a region of higher utility than what the best model allows to be achieved. The "blue" model actually has a smaller area under the ROC curve than the "yellow" model but the "blue" curve has a higher utility based on the net benefit of readmission-preventing actions based on the model's predictions. We see this because the blue ROC curve crosses over into the region of higher net utility on the graph.

In fact, several other models could be created during the process of learning that have even higher utility, but usually those will never be considered because the choice of the best model is often based on measures such as the area under the ROC curve rather than utility.

This example illustrates how a 2-step process to evaluate net benefit will fail to uncover that a model more useful than the best model, based on the area under the ROC curve, exists.

There are people who are interested solely in the accuracy or precision or AI models and they generally ignore the fact that their methods have no clinical utility. A decision curve analysis can help understand which model likely has the highest clinical utility - which is not always the model with the highest accuracy.

FEASIBILITY

For the feasibility of the algorithm, one must consider:

- Data availability and quality
- Implementation costs
- Deployment challenges
- Clinical uptake
- Maintenance over time

Data availability and quality. Data are an important aspect for any algorithm that learn from data. It is important when evaluating an AI model to look very closely at the data used for training, validation and testing.

- Includes data retrieval, preprocessing, and data cleaning
- Important to know if the population benefiting from the model is well represented in the training data
- Consider how the label values were assigned and who assigned these labeled values
- Transparency in the reporting of data and data processing is essential to evaluate any model

Other questions to think about regarding data:

- What is the ability of the data?
- Is it current and up to date?
- Are we using data that is five years old and trying to assess a new surgical procedure that was not well implemented in the time period of your dataset?

- How are missing data handled?
- Is the longitudinal data considered? If so, how is lost to follow up dealt with?

Feasibility of the action:

- Necessary resources. If you decide to act, do you have the necessary resources and equipment to perform that act?
- Necessary work capacity. Think about the work capacity necessary to act.

A component necessary to understand utility includes the Clinical Evaluation of the AI solution. The International Medical Device Regulator Forum (or the IMDRF) has developed a framework for clinical evaluation that was adopted globally, including by the US Food and Drug Administration (FDA). The framework is used to assess the risk and impact of AI solutions and to demonstrate assurance of safety, effectiveness, and performance.

Clinical evaluation:

1. Valid clinical association: Refers to the extent to which the model output is clinically accepted or well-founded based on an established scientific framework or body of evidence, and corresponds accurately in the real world to the healthcare situation and conditions identified by the AI solution. It is important to have a valid clinical association between your output and feature(s) if you expect clinical acceptance or clinical uptake.
2. Analytical validation: Does the model correctly process input data to generate reliable, accurate, and precise output data? This type of evaluation requires an understanding of the clinical data used to develop the model and the transparency in the reporting of the processing of training data, as well as a clear understanding of the labeled input and output variables.
3. Clinical validation: Measures the ability of an AI solution to generate a clinically meaningful output in the target disease, situation, or condition intended. Clinically meaningful refers to the impact the AI solution may have on the health of an individual or population.

Defining the accuracy and predictive value of an AI solution is needed, but the true evaluation of AI healthcare is not easy. Current AI in healthcare evaluation systems are limited - or non-existent - in their applicability for estimating the net utility of model. In addition, good examples of deployment AI solutions across systems are limited. As a result, healthcare teams have to rely on their personal experience and the collective experience of their colleagues to bridge the "gap" between available evidence and the needed evidence on AI evaluations.

CITATIONS AND ADDITIONAL READINGS

FDA, U. 2019. "Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)." FDA.

Shah NH, Milstein A, Bagley, PhD SC. Making Machine Learning Models Clinically Useful. JAMA. 2019;322(14):1351–1352. doi:10.1001/jama.2019.10306

MODULE 3: AI DEPLOYMENT

LEARNING OBJECTIVES

1. Understanding the four components of AI deployment
2. Describing the role of academics in the development and deployment of AI solutions in healthcare
3. Understanding the investments required to integrate AI solutions into the care setting, from the researcher to the clinician to the healthcare system
4. Knowing the major challenges, one needs to overcome to successfully deploy an AI solution into healthcare delivery, including consideration of foreseeable or intended harm

AI DEPLOYMENT

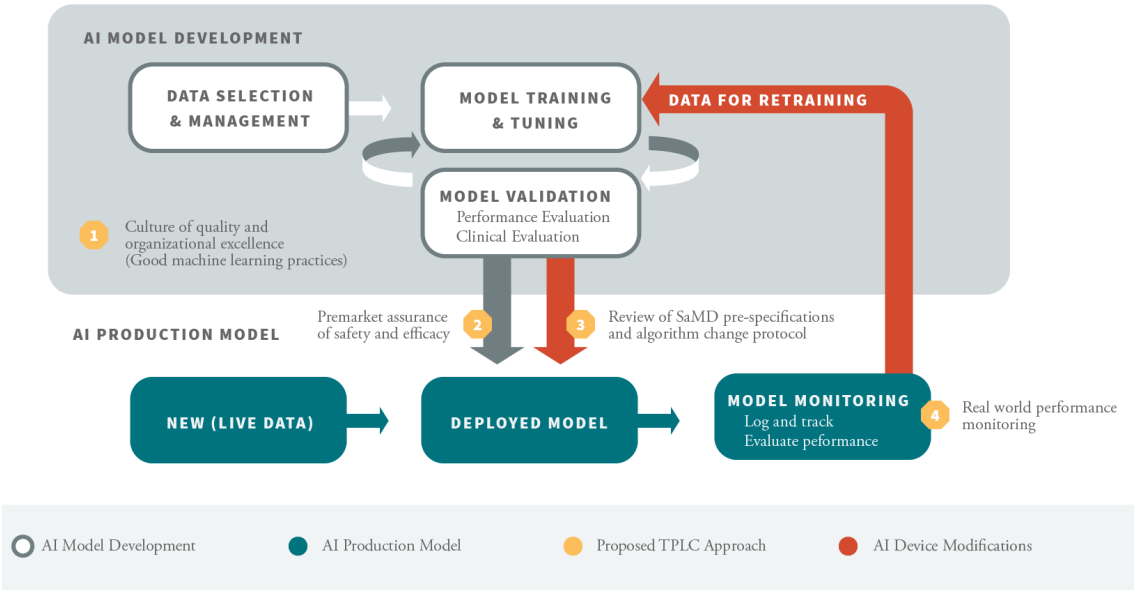
AI in healthcare has moved slower because human lives are at stake and therefore regulations, policies, and standards have set a higher bar for product development and deployment. Enthusiasm for AI in healthcare has been overshadowed by the challenging path to successfully implement these solutions into routine clinical care.

Challenges to deployment:

- Lack of buy-in from necessary stakeholders
- Lack of evidence for clinical utility
- Lack of resources and expertise in AI deployment
- The secondary use of clinical data that was generated and managed in the medical system with a primary purpose to support care and generate claims rather than facilitate clinical research and secondary analyses

To date, most literature and evaluations of AI solutions in healthcare have focused on proof of concept and the predictive abilities and accuracy of the AI solution. More recently, there has been an emerging focus on the downstream evaluations of AI applications in healthcare, including bias and fairness concerns across populations.

When we think about moving an AI solution into the healthcare setting, we need to think about a holistic framework for translation of the product into the delivery system - and multiple stakeholders interacting with the final product.



The AI model is one component of the **total product life cycle (TPLC)** and all components are necessary to understand and evaluate for deployment. The TPLC includes:

- Data Selection and Management
- Model Training and Tuning
- Model Validation

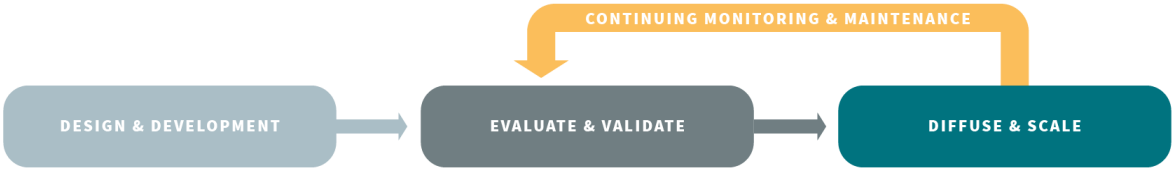
Once you have completed the first section "AI Model Development", you move to the AI Production Model, or the "Deployed Model" which continuously receives new (or live) data and is continuously monitored for safety and performance evaluations.

Some practical questions that you might (or should) ask prior to deployment:

1. **Research Question,** which addresses **Clinical Utility.** What is the clinical question and can I answer the question with an AI product? Is it a question that can be resolved with the data I have and the accuracy of the models I build?

2. **Stakeholder Involvement - Intended User.** Who will use the intended user of the model output and how would they want to see the results? Will the results from the model need to be displayed to the clinician at point of care, to the hospital management team to evaluate readmissions, or to an insurance company to identify high cost patients? Have all stakeholders involved evaluated both the input data and intended output data?
3. **Training Data.** What is the source data that I will use to train my models and where will I get this data? How recent are the data and what is the quality of data? Do your data need to be updated weekly, nightly, or hourly to make your prediction? What is the population distribution of this data and is the outcome of interest well represented in this data and does it represent the applied population?
4. **Implementation Costs - System Setup.** What is the system setup you will need to run the model? Is the system in which you developed your model interoperable with the system in which you will use your model? Will you be able to have all the necessary settings in the system you will use in the real application setting?
5. **Feasibility - Clinical Uptake.** How do you get the model output back into the workflow correctly and in real-time? Will your output be integrated into the EHR system or will you have to display the output alongside the EHR system? How will your model output interrupt the clinical workflow? Do you have significant buy-in from your end user and will they be your champion for deploying the system?
6. **Maintenance over time.** Who will be responsible for the continuous monitoring and updating of the AI product? How will new data be added or how and when will the model be re-trained? What happens if biases or unintended consequences are discovered? Whose responsibility will it be to update and maintain the product?

These questions must be evaluated prior to the deployment of the AI solution.



Deployment pathway has four fundamental components.

1. **Design and development of the AI product** - identifying the right problem to solve.
2. **Evaluation and validation** of the AI product, which will include multiple iterations. The model will likely first be evaluated using retrospective data
3. **Diffuse and scale** the AI production
4. **Continuous monitoring and maintenance** of the AI solution for safety and effectiveness.

DESIGN AND DEVELOPMENT

The design and development of the product begins with the problem: Is this an important problem and can it be answered with the data and AI model that is available? This step needs to be completed before deployment.

It is important to define and characterize the problem to be addressed by an AI solution and to evaluate whether that problem can be solved (or is worth solving) using AI; meaning identifying a model output that is actionable and will impact either care delivery or patient outcomes. It is important to understand the outcome you are defining in the context of the healthcare system.

It is also important to remember that medical definitions evolve over time. No medical definition is a static definition. When designing an AI solution, it is important to remember that the solution must be flexible, it must be able to accommodate a change in definition, a change in technology, a change in data type, or a change in the healthcare delivery system.

The health care setting is not a static field; clinical care guidelines and standards of care are constantly evolving. AI algorithm must be flexible to the point that it can adapt these new functionalities or changes over time with minimal effort.

Finally, it's important to remember that with any change, there will need to be new code reviews, new model performance metrics, and new monitoring capacities setup.

In the design and development component, **stakeholder involvement** must be assessed. Importantly, one must understand **the intended user**, their existing workflow, and how the output of the model will be delivered for action to this intended user.

Each different intended user would need a unique pathway or dashboard to receive the model output. For the respiratory distress example, who are your stakeholders?

- The IT team - you need to get the data in real time
- The Clinician - you need to know how and where they want to receive the data
- Organizational leaders - do you have the buy-in to deploy an AI product in the healthcare system
- Financial leaders - who is going to cover the overhead costs
- UI experts - who will you design the interface for your model output
- EHR systems experts - how will you integrate your results with the existing system
- Any many more users

Identifying the right stakeholders and having them involved in the design and development of AI products is a key to successful deployment.

Training Data - Data sources, data types and data availability are the crux of any AI product and knowing your data quality, reliability, representativeness and updates are imperative to AI deployment. AI products can utilize 3 different data sources:

1. Internal data sources
2. External data sources. This would be a team at one setting who is working with data from another setting. There is limited control on the quality and missingness of external data, unlike the internal data, unless one goes through a data agreement application - which can be laborious and difficult.
3. Public data sources. Many AI products are developed using public data sources, for example the MIMIC-III dataset, which is a freely accessible critical care database. An advantage of using public data sources is that the AI products developed and deployed with these datasets can be evaluated and reproduced by other scientists - increasing their reliability and impact in the field.

Whether AI products are developed from internal, external, or publicly available datasets, each data source has opportunities and challenges related to its use.

When designing and developing an AI product, it is important to **consider the setting and funding (Implementation Costs - System Setup)** of the original question, as this will drive or possibly hamper deployment. AI products that begin in industry are generally marketable and diffusible.

- Industry is often forced to partner with either academia or other industries in order to obtain the data needed for training and development of their models. Therefore, it is important to understand the data type and source from industry-lead products. Particularly their representativeness of the population to which the product will be applied.
- In academia, you have lead scientists who are thinking about innovative questions, innovative methodologies, and their focus usually begins with a specific research question. They have access to clinical data and clinical expertise. Given the collaborative academic environment, these teams are experts in developing multidisciplinary teams and these teams are easy to construct and maintain. However, academia has trouble recruiting and retaining

technical expertise - they generally cannot compete with industry salaries - and efficient scalability is generally not a strongpoint.

- AI products can also originate from the start-up setting. AI products from start-ups are usually self-funded, very efficient and focused on a single product or expertise. Often, when the startup companies have a good product, they are bought up by larger companies, who can then diffuse and scale their product.
- Philanthropic organizations generally work on target areas and partnerships. They might fund product development that addresses patient safety across settings. Sometimes philanthropy will partner with academics, industry, or government, as necessary, to develop their products of interest.

The design and development of the AI product can begin in many different settings, including industry, academics, start-ups, philanthropy, or government settings. Each setting has a unique set of challenges and opportunities for deployment.

EVALUATE AND VALIDATE

The second component of AI Deployment is the evaluation and validation of the AI product. Prior to deployment, the initial AI product must undergo rigorous in silico evaluations which includes:

- Investigation of the clinical utility (or net utility)
- Statistical validity
- Economic utility of the AI model

The utility of the AI solution (**clinical utility)** is likely the most important criteria to evaluate when considering the deployment of an AI solution in medical care. The utility of an AI solution relates to its applicability and impact on the healthcare system.

Factors that can affect clinical utility:

- Who needs to take the action
- Lead time offered by the prediction
- The existence of mitigating action (or therapy)
- The ability to intervene
- The logistics and cost of the intervention, incentives, etc.

To understand clinical utility for any AI solution you must ask: What is the primary task of the model and who are the main stakeholders, which is known as the outcome-action pair framework.

Clinical Utility is related to how well the AI product can demonstrate real-world workflow improvements or improve clinical care and patient outcomes. Clinical utility must be compared to baseline performance data - you have to show that the adoption of your product is useful, again in terms of clinical care (including clinical efficiency) and patient outcomes compared to current baseline performance data.

Net utility is related to the usefulness of the AI solution given the prevailing constraints in the care environment. Methods such as decision curve analyses can quantify the net benefit of using a model to guide subsequent actions given the costs of alternative actions, their corresponding benefits, and the various measures of model performance.

Net utility should be examined upfront, in order to have a useful model on the front line. One must consider the costs and benefits of the actions triggered by the AI product in order to form better decisions. The economic utility asks the question: Is there a real net benefit from the investment, or what is the cost of operational integration. For this, you must think about cost savings, increased reimbursement, or increased efficiency related to AI product.

Work capacity refers to the ability of a system to respond to a prediction. Work capacity is an important component of AI evaluation that needs to be evaluated prior to considering the deployment of an AI solution in healthcare. During this evaluation, it is also important to consider optimal utility (taking action on people who will benefit the most). Optimal utility is extremely important when predicting the use of scarce resources.

Net utility and work capacity are often ignored when AI products are reported in the scientific literature, yet they are essential to investigate prior to deploying an AI product in the healthcare setting.

Another aspect of the evaluate and validate deployment component is **statistical validity**. This includes performance metrics, such as accuracy, reliability, precision, recall and calibration. The statistical validity of a model is essential and often reported as a marker of model performance. However, there are a lack of guidelines for discrete levels of performance. The most accurate algorithm is often not necessarily the best algorithm to deploy. Statistical validity is **a component of deployment**. Identifying standards and markers of algorithm performance is becoming more

important and an emerging concept suggests there be a compliance or conformance component to the performance evaluation of AI products.

PRODUCT VALIDATION

When deploying, one must ascertain human engagement. Will humans be involved in the loop or will the AI product work autonomously and define actionable insights? This is one of the most important aspects in the evaluation and validation of an AI product before live integration into clinical care.

In **silent mode**, the AI product is deployed at point of care and predictions are made in real time but no action is taken on the predictions. The predictions are provided to the intended who then evaluates if the predictions are good or not or if they can be used to improve either the workflow or patient outcomes. This is crucial for finalizing workflows and product configurations, as well as the prospective, temporal validation of an AI product.

For care integration, an important step in this pathway is to consider the human-machine interaction:

1. Defining the clinical problem. It is important to identify a problem that is suitable to be addressed by the AI algorithm.
2. Think about the human-machine interaction.

The silent evaluation is very important to ensure the human - or intended user of the model output - is interpreting the model output correctly and the output is being applied appropriately and to the correct population. When we assess the human machine interaction, we need to think about how the clinical workflow is designed and how it will implement the AI product. In addition, you need to test the usability of the interface and the effect of your product on clinical decision making, including the legal and ethical issues of your AI product. Silent mode is an important, although often overlooked aspect of deployment.

There is an enormous gap between AI developed for research and AI deployed into clinical care settings. Therefore, **Clinical integration** might be the most difficult part of the deployment process. Some key considerations for the clinical integration of an AI product includes (1) Structural Considerations, and (2) Partnership Considerations.

Structural considerations:

1. Organizational capabilities
2. Personnel capacity
3. Cost, revenue and value considerations
    a. Initial costs
    b. Anticipated return on investment
    c. The value related to the AI deployment
4. Safety and efficacy surveillance
5. Cybersecurity and privacy

Partnership considerations:

1. Stakeholder consensus
2. Securing commitment from organizational leadership
3. Identifying leadership
4. Engaging stakeholders
5. Define milestones, metrics and outcomes to measure successful deployment

Clinical integration, while only a small mark on our pathway, is likely the biggest hurdle to overcome for a successful AI deployment.

Technology in the research environment greatly differs from the hospital environment. Significant effort and infrastructure investments are required to integrate AI products into real-time systems at point of care. One must consider the data platforms involved, the platform environment and the specific technology needed to get the data to run your models at point of care. Due to lack of interoperability and data standards, when another organization would like to implement a product already developed, they must also incur the same cost because they have to go through the same data gathering, cleaning, model evaluation and validation process as the original product development. Given the cost of real-world implementation of AI products, operational integration of the model should be considered carefully.

DIFFUSE AND SCALE

The third phase of deployment is to diffuse and scale the product. Diffuse and scaling the product comes after you solve local healthcare setting.

Three different types of systems to consider when developing deployment modalities:

- Fully integrated into the EHR system
- Partially integrated into the EHR system
- Standalone models

In order to diffuse the AI products, it must be able to ingest different data from different systems and support on premise and cloud deployment. A well-designed product would be able to adapt to an epic system or a Cerner system, or any other homegrown EHR system. Most modules up to date are stand alone.

It is important to understand that the majority of the products on the market, originally were developed in academics. Under the academic setting, products rarely get diffused and disseminated at scale, thus, they generally are coupled with industry partnerships to develop the full product. Products get externally licensed and scaled and diffused via commercial entities. Products are funded either through venture capitalists, or government in the healthcare set in academic setting.

MONITORING AND MAINTENANCE

Once an AI product is deployed a plan must exist to ensure the product will be continuously monitored and maintained. This will include regular architecture updates, addition of new training data, and perhaps yearly and/or irregular updates when industry reference files change.

Deterioration of model performance can occur within the same healthcare system over time when, for example, clinical care environments evolve, patient populations shift, or rates of exposure or outcomes change. A new code to diagnostic code for a disease of interest or a new clinical definition for an outcome of interest might become available. This would require an update of the AI product to account for these changes. There are also minor model updates or bug fixes that will need to happen at irregular time frames.

There are a number of approaches used to account for systematic changes to source data. These methods range from completely regenerating models on a periodic basis to recalibrating models using a variety of methods. However, the frequency and volume of these changes are not standardized.

Major and minor model improvements or new functionalities to address evolving clinical deeds are important.

CHALLENGES OF DEPLOYMENT

Deployment is complex and many issues need to be addressed before, during and after the implementation of an AI product in the healthcare system.

An important challenge in AI deployment is the **ethical challenge**:

- Data security and patient privacy: patients may be unaware their data are being used, shared, or sold for AI product development. In some healthcare settings there is a waiver of consent.
- Training samples not being representative of the intended population: This issue is further amplified because most AI products are not transparent about their training samples and often the demographic distribution of the training data is not reported.
- Transparency: The details about the evaluation metrics and validation are often not reported.
- Interoperability: If one system would like to deploy a product that was developed at another setting, they will likely need to re-incur the same cost as due to interoperability, most systems cannot seamlessly exchange code. New standards are emerging, such as SMART and FHIR.
    - FHIR is a standard for health care data exchange, published by HL7
    - The SMART App Framework connects third-party applications to EHR data, allowing apps to launch from inside or outside the user interface of an EHR system
- Lack of best practice standards for performance measures: There are no standard performance metrics for these models.
- Stealth science: Stealth science refers to science that is developed and disseminated without rigorous peer review. In industry, stealth science is common where companies may try to protect their trade secrets or avoid academic scrutiny. This is a particular for AI and healthcare, particularly as many AI products are developed by industry.

The models developed in research studies rarely translated into clinical care, hence, it is challenging to evaluate their real clinical and economical effect. **Prediction of sepsis** is a good example to go through to show how machine learning models for sepsis prediction can be translated into clinical care workflow:

- Sepsis Watch: The product was internally validated (prospectively) through a registered clinical trial and then licensed for commercial use in 2019
- Dascena Insight: The product was externally validated in a prospective clinical trial and retrospectively across 6 institutes to access generalizability.

- TREW Score: Developed using a publicly available dataset (MIMIC-II). TREW Score has been implemented in several hospitals.

There are several more similar algorithms; One might ask, why are there so many algorithms performing the same predictions and what is the best algorithm to deploy? It is important to think about all of the challenges we have discussed regarding deployment and think about how one can evaluate or compare these like AI products.

CITATIONS AND ADDITIONAL READINGS

Gupta, A., T. Liu, and S. Shepherd. 2020. "Clinical decision support system to assess the risk of sepsis using Tree Augmented Bayesian networks and electronic medical record data." *Health Informatics J* 26(2): 841-61.

Sendak, M. P., W. Ratliff, D. Sarro, E. Alderton, J. Futoma, M. Gao, M. Nichols, M. Revoir, F. Yashar, C. Miller, K. Kester, S. Sandhu, K. Corey, N. Brajer, C. Tan, A. Lin, T. Brown, S. Engelbosch, K. Anstrom, M. C. Elish, K. Heller, R. Donohoe, J. Theiling, E. Poon, S. Balu, A. Bedoya, and C. O'Brien. 2020. "Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study." *JMIR Med Inform* 8(7): e15182.

Sendak, M.P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., Ratliff, W. and Balu, S., 2020. A path for translation of machine learning products into healthcare delivery. European Medical Journal Innovations.

Shah NH, Milstein A, Bagley, PhD SC. Making Machine Learning Models Clinically Useful. JAMA. 2019;322(14):1351–1352. doi:10.1001/jama.2019.10306

Topiwala, R., K. Patel, J. Twigg, J. Rhule, and B. Meisenberg. 2019. "Retrospective Observational Study of the Clinical Performance Characteristics of a Machine Learning Approach to Early Sepsis Identification." *Crit Care Explor* 1(9): e0046.

MODULE 4: DOWNSTREAM EVALUATIONS OF AI IN HEALTHCARE: BIAS AND FAIRNESS

LEARNING OBJECTIVES

1. Overview of Bias and Fairness in AI solutions in healthcare
2. Understand the different types of bias in AI healthcare solutions
3. Describe algorithmic fairness
4. Identify solutions to address bias and fairness in AI solutions

BIAS IN AI SOLUTIONS

Recently, reports have questioned whether AI solutions in healthcare might actually perpetuate discrimination if trained on historical data—which are often poorly representative of broader populations. Often, AI models are trained using historical or retrospective data which are often derived from affluent academic medical centers that likely do not contain all populations, particularly diverse populations for which the AI solutions will be applied. AI models exclusively trained on such data may further perpetuate disparities and fail to demonstrate external validity in broader patient communities. This is likely due to a lack of diversity represented in the training data, a lack of understanding how a disease may manifest and progress in different populations, and a lack of human understanding of the potential consequences and biases that may be inherent in AI solutions.

Fairness and bias in AI solutions may be a larger problem in countries where important health disparities exist based on patient demographics, such as the United States. Therefore, as tools proliferate across clinical settings, it is important to think about and understand potential demographic biases underlying model development and deployment.

Heart failure example:

> Not long ago we didn't know that symptoms of heart attack look different in women compared to men, which led to differences in cardiovascular mortality rates between women and men. The problem here is that the model was developed based on male symptoms so the model might be very accurate for identifying males with heart attack symptoms, but it might not work well in females, who present with different symptoms. The predictive accuracy, when analyzed against the true clinical outcomes, will decline for women, but not men.

Genomic database example:

> Genomic databases used across the research community where we find major racial bias in the genomic samples collected. These databases are the center of precision medicine, where our ability to identify whether a genetic variant is responsible for a given disease or phenotypic trait depends in part on the confidence in labelling a variant as pathogenic. Research suggests that these databases heavily reflect European ancestry, and in fact are missing major population-specific pathogenic information, particularly African-specific pathogenicity data. Therefore, genomic test results for persons of non-European ancestry could be less accurate, more challenging, or simply unattainable.

Dermatology example:

> In general, patients with darker skin present with more-advanced skin disease and have lower survival rates than fair-skinned patients. It is possible that the only fair-skinned populations may benefit because of the lack of inclusion of darker skinned patients in model training and development. If the algorithm is basing most of its knowledge on how skin lesions appear on fair skin, then theoretically, lesions on patients of color are less likely to be diagnosed and therefore benefit from the AI solution.

These examples provide you with an idea of how wide-spread the challenges are related to fairness and bias in AI solutions for healthcare.

TYPES OF BIAS

Bias can occur during almost any stage of AI model building and implementation - from data collection to model deployment.

Types of bias:

- Historical bias
- Representation bias
- Measurement bias
- Aggregation bias
- Evaluation bias
- Deployment bias

Historical bias occurs if the present or past state of the world influences a model in a way such that its predictions are considered unfair given societal values and norms. Historical bias refers to judgement based on preconceived notions or prejudices. AI algorithms are entirely data dependent and historical bias encoded in real-world data cannot even be overcome by perfect sampling and feature selection. Therefore, it is important to remember that historical healthcare data, in general, is extremely male and extremely white, and this has real-world impacts.

Representation bias (also called sampling bias) arises when the sample collected to train an AI solution does not represent the actual distribution of the population it is intended to be applied to. It occurs when certain parts of the final use population are underrepresented in the training data.

Measurement bias arises in situations if the noise is not randomly distributed but differs across groups, which leads to differential performance. Often the only available and measurable features as well as labels are only noisy proxies of the actual variable of interest. Usually, one cannot change the data, some historical biases might be indistinguishably linked to the data – but the awareness about the problem is important and mitigation strategies can be identified by taking preventive measures such as pre- and post-processing actions.

Aggregation bias occurs while developing the model when we try to combine different populations whose underlying distribution of the outcome under study differs. This problem is known as infra-marginality and requires separate models for the different populations or including the demographic variable into the model to account for the systemic differences. In terms of model development, one size does not fit all. In order to identify aggregation bias, developers need to understand the meaningful distinct groups and reasons why they are different from each other.

Evaluation bias occurs during the model validation and tuning. Evaluation bias arises if the testing data, which often includes external benchmark datasets, is not representative of the final population to which the AI solutions will be applied. This is the difference between the data used for model evaluation and the data used for model's real-world predictions. Since developers mostly use a benchmark dataset or a synthetic dataset for training, their evaluation often does not fit the real-world. A solution of this problem is external validation of the AI model on a different unseen data selected from the targeted population. Evaluation bias can arise if inappropriate performance metrics are used. Evaluation bias also refers to usage of evaluation metrics inefficiently and to avoid it, using granular and comprehensive evaluation metrics is suggested.

Deployment bias arises during the implementation of the model. It refers to using the model inappropriately or misinterpreting its results. In other words, if the model's intended use is different from the way it is used, deployment bias occurs. Deployment bias is the interaction of society with the AI solution - how society or the medical community uses the AI solution and its output.

ALGORITHMIC FAIRNESS

Ethical analysis of AI Solutions in healthcare demand that we take a view of fairness, or more appropriately, justice that centers on the health and lives of people, not the outputs alone. A lot of historical medicine has been influenced by white normativity, which is the basis of many medical facts. This is evident because of a lack of inclusion of diverse patients in clinical research. have knowledge that the insiders don't. Bringing diverse perspectives actually enhances the quality and accuracy of your scientific endeavor.

A key difficulty in developing fair AI algorithms is the fact that no universal notion of fairness exists. Many different definitions have been proposed by researchers over the years and they can be broadly regrouped into three main classes: **anti-classification, classification parity** and **calibration**. These fairness definitions have been shown to all suffer from significant statistical shortcomings. Therefore, special caution and awareness about the notion's limitations and weaknesses is always necessary when applying these concepts in model evaluation settings.

Anti-classification or **"fairness through unawareness"**

1. Requires the exclusion of any protected attributes in the outcome modeling
2. Requires the omission of any unprotected characteristics that are proxies of protected attributes

The main shortcoming of this fairness definition is that some clinical risk models need to explicitly include protected attributes for it to be equitable. In particular, this applies to situations where the true underlying risk distribution differs across subpopulations, known as **the problem of infra-marginality**. Therefore, an accurate model must include protected attributes, but must also learn to avoid bias based on these attributes.

AI solutions that use datasets which may be under-representative of certain groups, may need additional training data to improve accuracy in the decision-making and reduce unfair results. Anti-classification requires the exclusion of any protected attributes in the outcome modeling.

Classification parity of fairness asks for equal predictive performance across any protected group. When selecting the metrics to examine, it is important to keep in mind the actionable insights resulting from a model output.

Two measures have received particularly high attention by the machine learning community:

1. False positive rate: The probability of predicting a positive outcome when the true outcome is negative should be the same regardless of protected attributes of the patients
2. Proportion of positive decisions: The probability of predicting a positive outcome should not vary across different demographic groups given all else equal. It is also known as demographic parity as it requires the classifier's predictions to be independent of protected traits.

These definitions are problematic when the risk distributions are different for different groups, a problem known as infra-marginality. Classification parity asks for equal predictive performance across groups.

Calibration is when a model reaches a good agreement between model predictions and observed outcomes. Calibration means that when conditioning on risk estimates, outcomes should be independent of the protected attributes. Think of calibration as a comparison of the actual output and the expected output given by a system. Calibration is open to manipulation of risk distributions for different groups. Model calibration is an important aspect of development and must be evaluated before model deployment.

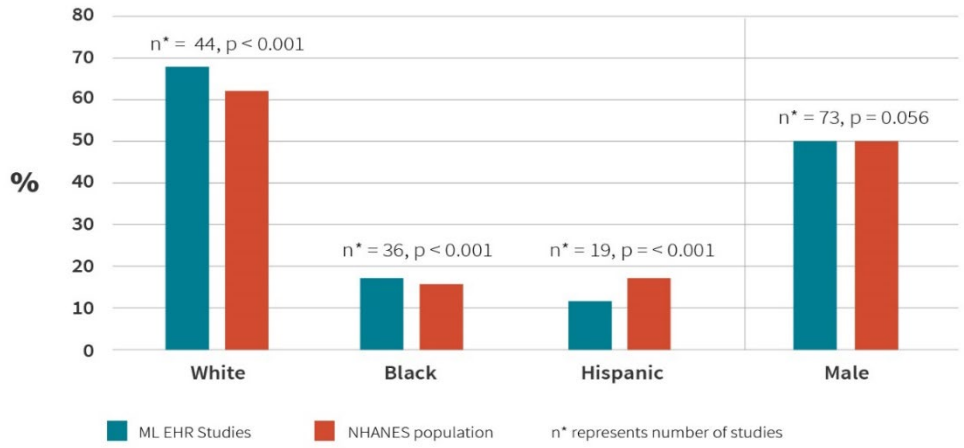
Applying fairness measure:

- Anti-classification: Checking the requirement of not using any protected attributes in the decision rule. It gets more complicated when trying to impose the stricter condition to also not use any proxies of protected attributes.
- Classification parity: Once the 2-3 most relevant performance metrics are selected, we evaluate performance differences on the test-set across different demographic groups. It is always better to calculate empirical confidence intervals of this statistic to decrease the dependence on the test set.
- Calibration: Only look at the overall rate of predicted and observed outcomes across demographic groups.

TRANSPARENCY

It is important to think about whether the population captured in the EHRs system is representative of the broader community, particularly if these are coming for an academic medical center. If AI solutions are being developed on non-representative populations, the utility and applicability of these advances for the broader patient community falls into question. This relates to an important

issue around **transparency and reporting** and many AI studies, particularly machine learning, do not report the demographic breakdown of the training data used to develop and train the models. Lack of reporting makes it very difficult to evaluate the bias and fairness of the AI solution and its applicability across populations.



Studies that mentioned variables often did not report if they were included as model inputs. In the studies that reported demographics, the average populations included higher proportions of whites and Blacks yet fewer Hispanics compared to the general population. While each study might not be applicable to the general population, these findings emphasize the present lack of transparency in reporting details of training data used for development and evaluation by machine learning models in healthcare. To ensure the unbiased deployment and application of any AI model in healthcare, detailed information on the data used to develop and train the model are necessary.

As a solution for transparent reporting and to identify best-practices for designing machine learning models to account for biases and fairness, MINIMAR template is suggested. (MINIMAR = The **MIN**imum **I**nformation for **M**edical **AI** **R**eporting)

MINIMAR Requirements:

1. Include information on the population providing the training data, in terms of data sources and cohort selection
2. Include information on the training data demographics in a way that enables a comparison with the population the model is to be applied to
3. Provide detailed information about the model architecture and development so as to interpret the intent of the model and compare it to similar models

4. Model evaluation, optimization, and validation should be transparently reported to clarify how local model optimization can be achieved and to enable replication and resource sharing

You can understand that by providing these details of the training data and population, model design and intent, an end-user will have a great understanding of how to best deploy the model and in which populations.

DOWNSTREAM EVALUATIONS

While we have covered many topics in this lecture regarding bias and algorithmic fairness, there are still many more challenges and opportunities for Fair AI research:

- Defining fairness: There are several definitions of AI fairness that have been proposed in the literature. It is nearly impossible to understand how one fairness solution could address all challenges. Identifying the correct or best definition is an ongoing debate.
- From equality to equity: Equity suggests that each group is given the amount of resources needed to have similar outcomes. Understanding how to develop and implement a model that provides both equality and equity presents a paradigm shift in the way to think about healthcare delivery and is an active area of research.
- Identifying biases in models, and particularly in datasets: Many biases are systematic and we are often unaware they exist. We still have a long way to go before we can systematically mitigate these biases and provide our professionals with the appropriate tools they need to address these issues at point of care.

The perspective collection and reporting of AI outputs, clinical recommendations and patients decisions coupled with eventual outcomes is essential in being accountable as healthcare institutions and as clinicians. This work and transparency in reporting AI solutions is absolutely critical for populations who have difficulty trusting the medical establishment. The key is to use AI in a way that actually does benefit all groups, which requires thoughtful evaluations and human interpretations.

CITATIONS AND ADDITIONAL READINGS

Adamson, A. S. and A. Smith. 2018. “Machine Learning and Health Care Disparities in Dermatology.” *JAMA Dermatol* 154(11): 1247-48.

Beery, T. A. 1995. "Gender bias in the diagnosis and treatment of coronary artery disease." *Heart Lung* 24(6): 427-35.

Bozkurt, S., E. Cahan, M. G. Seneviratne, R. Sun, J. A. Lossio-Ventura, J. P. A. Ioannidis, and T. Hernandez-Boussard. 2020. "Reporting of demographic data, representativeness and transparency in machine learning models using electronic health records."

Corbett-Davies, S. and Goel, S., 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Hernandez-Boussard, T., S. Bozkurt, J. P. A. Ioannidis, and N. H. Shah. 2020. "MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care." *J Am Med Inform Assoc*.

Kessler, M. D., L. Yerges-Armstrong, M. A. Taub, A. C. Shetty, K. Maloney, L. J. B. Jeng, I. Ruczinski, A. M. Levin, L. K. Williams, T. H. Beaty, R. A. Mathias, K. C. Barnes, T. D. O'Connor, and C. o. A. a. A.-a. P. i. t. A. (CAAPA). 2016. "Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry." *Nat Commun* 7: 12521.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366(6464): 447-53.

Parthipan, A., I. Banerjee, K. Humphreys, S. M. Asch, C. Curtin, I. Carroll, and T. Hernandez-Boussard. 2019. "Predicting inadequate postoperative pain management in depressed patients: A machine learning approach." *PLoS One* 14(2): e0210575.

Spanakis, E. K. and S. H. Golden. 2013. "Race/ethnic difference in diabetes and diabetic complications." *Curr Diab Rep* 13(6): 814-23.

Suresh, H. and J. V. Guttag. 2019. "A framework for understanding unintended consequences of machine learning." *arXiv preprint arXiv:1901.10002*.

MODULE 5: THE REGULATORY ENVIRONMENT FOR AI IN HEALTHCARE

LEARNING OBJECTIVES

- Understand why most AI solutions in healthcare have not received regulatory approval, to date
- Describe best practices for AI development, in particular good machine learning practices
- Recognize the risk framework used to classify AI solutions in healthcare that is used by the Food and Drug Administration (FDA)
- Know the 3 concepts of model properties that can be regulated
- Understand main differences between EU, China and US regulations on AI solutions

OVERVIEW

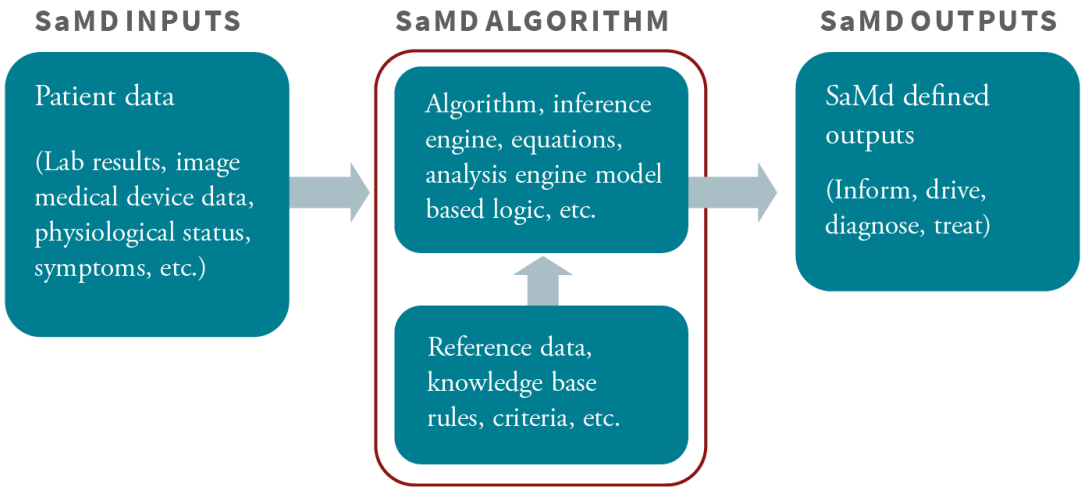
International Medical Device Regulators Forum (IMDRF): An international workgroup composed of AI regulators that come together to develop a path for standardized AI regulations.

Software as a Medical Device (SaMD): Software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device.

- Medical purpose: Intended to treat, diagnose, cure, mitigate, or prevent disease or other conditions
- SaMDs are NOT part of hardware

COMPONENTS OF REGULATION

The SaMD can be described in 3 components: SaMD inputs, SaMD algorithms, and SaMD outputs.



This is the framework used by the US Food and Drug Administration to regulate AI Solutions / SaMD.

The FDA's regulatory framework starts with a Market Application, which includes a definition statement and category (I, II, III, or IV). The category is based on the risk associated with the use of the proposed AI solution.

Depending on the category defined, data requirements necessary for regulation may include:

- A premarket notification (or 510(k) statement of equivalency)
- De Novo request (no similar product exists on the market for comparison)
- Premarket Application (PMA), which is reserved for high risk AI applications

Definition Statement is required for every SaMD application and is used to identify the submission category, which defines risk and subsequent data requirements.

The definition statement must:

- Clearly identify the intended medical purpose of the model (treat, diagnose, drive clinical management, inform clinical management)
- State the healthcare situation or condition that the AI model is intended for, which includes critical, serious, and non-serious conditions
- Include the intended population for the application
- Identify the intended users (or stakeholders) of the model

Using information from the definition statement, the SaMD Category is defined, which is based on a risk framework developed by the IMDRF.

STATE OF HEALTHCARE SITUATION OR CONDITION	SIGNIFICANCE OF INFORMATION PROVIDED BY SaMD TO THE HEALTHCARE DECISION		
	Treat or Diagnose	Drive Clinical Management	Inform Clinical Management
Critical	IV	III	II
Serious	III	II	I
Non-Serious	II	I	I

SaMD N12[2] Framework

In the framework, risk is set by the intended use of the SaMD and the state of the health situation it targets.

- The columns in the grid represent the **Significance of information provided for a healthcare decision**. This is the **ACTION** in the **outcome-action pairing** framework used for AI evaluation.
- The rows represent the **State of healthcare situation or condition**, which identifies the state of the healthcare situation or condition as critical, serious, or non-serious

An example that demonstrates the definition statement and risk category:
A SaMD AI application that "**drives clinical management**" in a "**critical healthcare situation or condition**" (Category 3):

- In a hypothetical situation, an AI application is developed for ICU patients that receives electrocardiogram, blood pressure, and temperature signals from a patient monitoring system. The physiologic signals are processed and analyzed to detect patterns that occur at the onset of patient instability and deterioration, a threshold set by a utility analysis. When physiologic instability is detected, an alarm is generated to indicate that immediate clinical action is needed to prevent potential harm to the patient.

SaMD regulations place devices into four categories based on the risks associated with the use of the device. Category I being the lowest risk; Category IV being the highest risk.

- Category I devices require general controls
- Category II devices are medium to moderate risk and require the same general controls and some additional controls
- Category III and IV applications require "general controls" and pre-market approval which is the most stringent regulatory category. To date, there are very few Category IV SaMDs that

are approved in the US. These are high risk, generally life-supporting, life-sustaining AI applications.

All applications must include general controls. General controls require that all AI solutions comply with three components:

1. Quality system regulations
2. Current good manufacturing practices
3. Properly labeled - the label of the device follows FDA guidelines and regulations

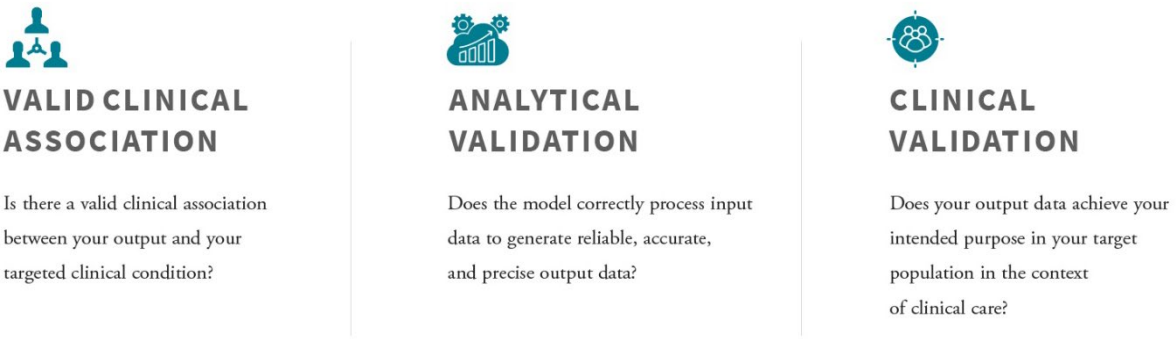
Any adverse event associated with the AI solution must be reported.

Quality System Regulations: Manufacturers are required to have processes in place for controlled bug resolution, incident reporting, standardized design processes and overall risk management.

CLINICAL EVALUATION PROCESS

All AI applications needing regulatory approval must include general controls. As part of the **general control**, **the Clinical Evaluation Process** is a framework used by regulators to understand quality system regulations. The IMDRF defines the clinical evaluation process as **ongoing activities** conducted for the assessment and analysis of a SaMD's clinical safety, effectiveness and performance.

The Clinical Evaluation Process includes three components.



Valid Clinical Association (Category I): Is there a valid clinical association between your SaMD output and your SaMD's targeted clinical condition? Scientific validity of the AI solution or the extent to which the SaMD's output is clinically accepted or well-founded (based on an established scientific evidence), and accurately corresponds to the healthcare situation and condition identified

in a real-world setting. An indicator of the level of clinical acceptance and confidence one can have of the SaMD's output

Minimum evidence to support the clinical association could include:

- Literature searches
- Original clinical research
- Professional society guidelines
- Examples of how your model can generate new evidence
- Secondary data analysis
- The performance of a clinical trial based on your AI solution
- Required for AI regulation and ensures the clinical acceptance and uptake of the AI solution in the healthcare setting

Novel associations - associations that are newly discovered by your AI:

- As literature and existing randomized clinical trials do not exist for this association, there are other solutions to regulate this software, which may include performing a clinical trial or secondary data analyses

Analytical Validation (Category II): Evaluates whether your AI solution correctly processes input data to generate accurate, reliable, and precise output data. Part of the verification and validation phase that should be performed by the AI manufacturer. Provides objective evidence that the AI solution was correctly constructed and the data processing is reliable.

May come as part of your good software engineering practices or from generating new evidence through use of curated databases or previously collected patient data

Clinical Validation (Category III): Does the use of your AI's output data achieve your intended purpose in your target population in the context of clinical care? Related to the positive impact of an AI Solution on the health of an individual or population.

Clinical Validation should be:

- Measurable, patient-relevant clinical outcome(s)
- Including outcome(s) related to the function of the model
- A positive impact on an individual or public health

Prior to product launch of the AI product (pre-market), evidence must exist on the following:

- AI accuracy
- Specificity
- Sensitivity
- Reliability
- Usability
- Limitations
- Scope of use in the intended use environment with the intended user

After launching the product (post-market) the product must:

- Continue to collect real world performance data to
- Further understand the healthcare needs to ensure the AI solution is meeting those needs
- Monitor the product's continued safety, effectiveness and performance in real-world use

The IMDRF identifies that clinical validation is a necessary component of regulation and that it can be demonstrated through several paths, including:

- Referencing existing data from studies conducted for the same intended use
- Referencing existing data from studies conducted for a different intended use, where extrapolation of such data can be justified
- Generating new clinical data for a specific intended use

The Clinical Evaluation Process is an important component of the regulatory environment.

FDA APPLICATION

In addition to the **general control process**, there are other regulatory control requirements (data requirements) that accompany an SaMD application that depend on the application's **category of risk (1 - 4)**, which can include one of these regulatory components:

- Premarket Notification 510(k)
- De Novo Notification
- Premarket Approval (PMA)

Premarket Notification 510(k) is the simplest data requirement, if the SaMD is similar to a product already on the market. The intent is to inform the regulatory agencies that the AI solution is safe and effective, which is determined by demonstrating the AI solution is equivalent to a legally marketed device, often known as a "predicate".

To determine equivalence, the AI solution must have:

- The same intended use as the predicate AND have the same technological characteristics, OR
- The same intended use as the predicate and a different technology that will not raise safety or efficacy questions AND the AI solution is at least as safe and effective as the predicate

As more and more applications become approved, pre-market notifications will become an easier and efficient pathway towards regulation.

De Novo Notifications: Submitted when there is no "predicate". Limited to Category I and Category II SaMDs and are a risk-based classification process.

The de Novo notification should include:

- Clinical data (if applicable) that are relevant to support the assurance of the safety and effectiveness of the AI solution
- Non-clinical data including bench performance testing
- Description of the probable benefits of the AI solution when compared to the probable or anticipated risks when it is used as intended

Premarket Approval (PMA): Required for high risk SaMDs (Category III and IV). The most stringently regulated application required by the FDA. Includes rigorous technical studies, non-clinical laboratory studies, laboratory studies, and clinical investigations.

Before PMA approval, the applicant must provide valid scientific evidence demonstrating reasonable assurances of safety and effectiveness for the device's intended use.

After an AI solution receives regulatory approval, a modification may be required in certain circumstances. A modification request must be submitted if there is new risk or a change of an existing risk.

A modification may be required if there is:

- A change to risk controls to prevent significant harm
- A change that significantly affects clinical functionality or performance. A change in clinical functionality could include
    - New indications for use
    - New clinical effects
    - Significant technology modifications that affect performance characteristics

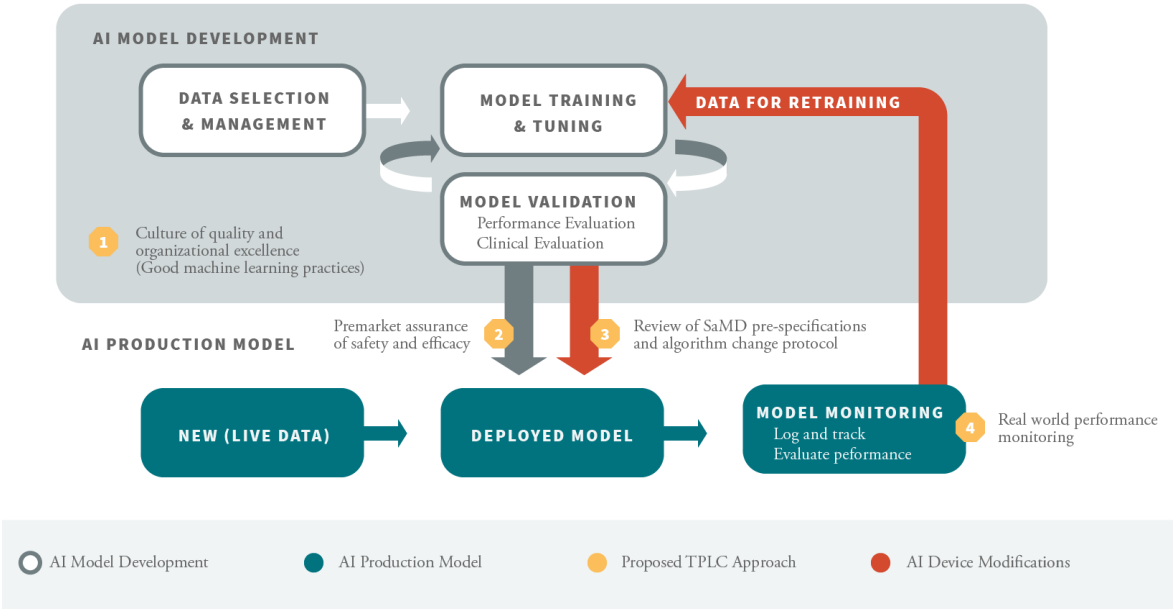
SaMD modifications generally fall into three categories:

1. Change in performance
    - Example: The incorporation of new training data or a change in AI architecture which could alter performance
2. Change to the model Input
    - Example: The incorporation of different sources of the same input or adding new inputs that were not previously considered
3. Change of the intended use of the output
    - Example: Change in disease (apply to new condition)

Software modifications are common and essential in the total life cycle of the AI Solution.

PRODUCT APPROVAL

In line with the framework proposed by IMDRF, the FDA has developed the following diagram related to the **total product life-cycle (TPLC)** for an AI workflow.

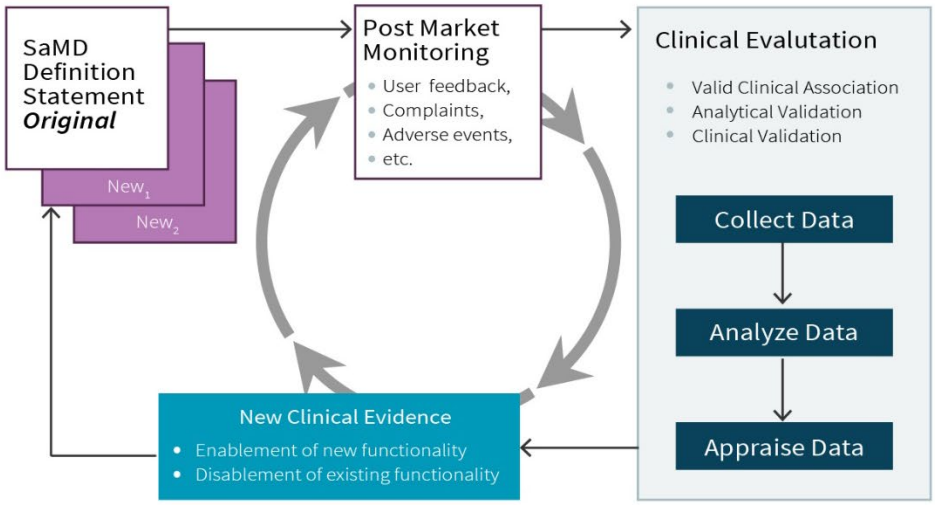


There are 4 distinct components in the total product life-cycle:

1. The culture of quality and organizational excellence, which is also referred to as Good Machine Learning Practices. This includes all components that must be considered when developing an AI solution, such as
    - Data selection and management
    - Model training and tuning
    - Model validation
2. Premarket assurance of safety and efficacy of the AI solution. It is expected that safety and effectiveness are continually monitored throughout the life cycle, including patient risks and patient safety. Regulators expect a manufacturer to perform a risk assessment and evaluate that the risks are reasonably mitigated throughout the TPLC.
3. Regular monitoring of safety and intended use, which is used to identify when a software modification is required. The regular monitoring of the deployed model is necessary and should include the ability to log and track model performance.
4. The Continuous Learning expected from an SaDM that Leverages Real World Data. "Continuous learning" is not "machine learning." It refers to collecting post-market information to update and evaluate your existing AI solution.

The TPLC diagram demonstrates how regulators are thinking about using real world data for continuous learning, the basis of the learning healthcare system.

Locked algorithms are AI solutions that provide the same results each time the same input is provided. Generally fixed functions to a given set of inputs, and may use a manual process for updates and validation.



Pathway for Continuous Learning – Use of real World SaMD Performance Data in Ongoing SaMD Clinical Evaluation

How to regulate "adaptive" AI solutions or machine learning applications?

- Following deployment, these types of adaptive models may provide a different output in comparison to the output initially approved for a given set of inputs. These automated changes use a well-defined process, which aim to improve outcomes based on new data or additional data that is taken as an input. There are two stages to an adaptive algorithm:
    - Learning stage: The algorithm "learns" how to change its behavior, based on the addition of new input types or new cases to an already existing training set
    - Update stage: The algorithm will update when the new version of the algorithm is deployed

This is a paradigm shift in the regulatory process and requires a new total product lifecycle that allows these devices to continually improve while providing effective safeguards.

Considerations for the regulatory environment for an AI model:

1. Developers should be aware of SaMD Risk Classifications, total product life-cycle (TPLC) and Good Machine Learning Practices (GMLP)

2. The population, performance and intended use are aspects of the model that can be regulated and any change to these characteristics after approval would require a notice of modification
3. The intended use and state of the healthcare situation of the AI model drives the level of regulation and risk category
4. Safety and effectiveness must be continuously monitored post-market using real-world data to ensure a learning health system

Example - Arterys

- Indications for Use: Arterys is a software that uses cardiovascular images acquired from magnetic resonance (MR) scanners. It analyzes blood flow from the heart using the MR images. The output is intended to be used to support cardiologists, radiologists, and other healthcare professionals for clinical decision making.
- Risk Classification: Class II
    - Significance of information is to "inform clinical management"
    - State of healthcare situation or condition is "critical"

Example - IDx-DR

- Indications for Use: IDx-DR is a retinal diagnostic software device that incorporates an adaptive algorithm to evaluate ophthalmic images for diagnostic screening to identify retinal diseases or conditions
- Risk Classification: Class II
    - Significance of information is to "drive clinical management"
    - State of healthcare situation or condition is "serious"

Example - Guardian Connect (Medtronic)

- Indications for Use: The Guardian Connect system is indicated for continuous or periodic monitoring of glucose levels in the fluid under the skin, in patients (14 to 75 years of age) diagnosed with diabetes.
- Risk Classification: Class II
    - Significance of information is to "drive clinical management
    - State of healthcare situation or condition is "serious"

The FDA does not regulate certain types of clinical decision support (CDS) tools under 21st century cures act.

Three criteria determine whether CDS are regulated as an SaMD:

1. The software cannot receive, analyze or otherwise process a medical image or signal from an in vitro diagnostic device or from any other signal acquisition system
2. A healthcare professional must be able to understand the basis of its recommendations
3. The software cannot be intended as the sole source of recommendations regarding treatment, diagnosis or prevention of a disease

The FDA doesn't regulate AI solutions that are "laboratory-developed tests" designed, developed and deployed within a single health care setting.

FDA's Digital Health Software Precertification Program (Pre-Cert):

- Streamlines regulation of AI solutions
- Organizations may become "approved" which would allow them to bring products to market without a premarket review, provided the product is "lower risk"
- Once certified, organizations can make certain minor changes to its AI products without having to submit a modification request

An organization must demonstrate the FDA's five quality and organizational excellence principles in order to be considered for the Pre-Cert program:

1. Product quality
2. Patient safety
3. Clinical responsibility
4. Cybersecurity responsibility
5. Proactive culture

GLOBAL ENVIRONMENT

It is important to note that there are some differences across the globe regarding regulatory guidelines for AI in healthcare.

EU's General Data Protection and Regulation (GDPR)

- Outlines a comprehensive set of regulations for the collection, storage, and use of personal information which may be used in AI solutions
- Describes the right of citizens to receive an explanation for algorithmic decisions. The implications would exclude the use of many types of algorithms used today in advertising and social networks - and eventually healthcare.

- Provides protection of data from EU citizens but given the global Al momentum, the laws may impact companies from the US and worldwide
- A critical component of this regulation is Article 22: "Automated individual decision making, including profiling." Article 22 requires explicit and informed consent before any collection of personal data.
- The "right to explanation"
    - Requires that meaningful information about the AI logic as well as the potential significance and consequences of the data-driven system are provided upon request
    - Refers to the use of the black box algorithms
    - Could potentially limit the types of models that manufacturers are able to use in health-related applications
    - Holds the manufacturers of AI-based technologies more accountable

While there are some differences between the US and EU regulations, a common theme from both entities is the protection of the individual, their data, and their right to information.

China leads in the number of AI patents as a result of this favorable environment. AI governance in China is aimed at the development of "responsible AI" and focused on the societal beneficiary rather than the individual beneficiary. China has also put a focus on regulating education so that the nation produces more STEM workers.

China requires businesses and private citizens to share their data with the government – almost the opposite of US and EU regulations. The incentives for data sharing and elimination of data silos could catapult China in clinically meaningful AI technologies.

Principles for AI governance released by China's Ministry of Science and Technology (MOST) include:

1. Harmony and friendliness
2. Fairness and justice
3. Inclusivity and sharing
4. Respect privacy
5. Secure/safe and controllable
6. Shared responsibility
7. Open collaboration
8. Agile governance

China is rapidly developing important AI solutions in healthcare and these governing policies are essential to balance both the innovation of technology as well as the safety of healthcare delivery.

The White House Office of Management and Budget (OMB) document provides Guidance for Regulation of AI Applications. The regulations are not specific to healthcare, but they provide the umbrella of regulations applied to AI solutions.

The ten guiding principles are aligned with the FDA's regulatory processes:

1. There must be public trust in AI
    - Public trust and validation is critical to the adoption and acceptance of these technologies in the healthcare sector
    - Privacy and other risks must be addressed with appropriate mitigation strategies that are well documented and transparently reported
2. There must be public participation in AI development
    - There should be ample opportunities for the public to provide information and participate at all stages possible of the "rule-making" process
3. AI solutions must be based on scientific integrity - clinical validation in the clinical evaluation process proposed by the IMDRF
    - AI in healthcare should be based on scientific and technical information and processes that are likely to have a clear and substantial influence on public policy or private sector decisions - and these standards should be held to the highest level of quality, transparency, and compliance
    - Best practices include clearly stating the strengths, weaknesses, intended optimizations or outcomes, bias mitigation, and appropriate uses of the AI application's results
    - Data used to train the AI system must be of sufficient quality for the intended use
4. Every AI solution must have a Risk Assessment and Management component
    - A risk-based approach should be used to determine which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits
    - If an AI tool fails, the magnitude and nature of the consequences will inform the level and type of regulatory effort that is appropriate to identify and mitigate risks
5. Benefits and Cost
    - For an AI algorithm to be deployed, it must offer significant potential benefit

   - Before implementing an AI solution, agencies must consider the full societal costs, benefits, and effects related to the development and deployment of these applications
6. AI solutions must be flexible
   - Performance based and flexible approaches should be easily adapted and updated - this would include the continuous monitoring of real world evidence to improve upon the AI solution
7. AI solutions must be fair and non-discriminatory
   - Transparency regarding potential biases and discriminatory aspects of the algorithm is becoming more and more important as we see potential harm due to biases promoted or exacerbated through AI
8. AI solutions that provide Disclosure and Transparency will promote public trust
   - Healthcare systems should disclose when AI solutions are in use and how these applications can impact patients and decisions
9. All AI solutions must address Safety and Security issues
   - Safety and security should be considered throughout the design, development, deployment, and operation process
   - There should be controls in place to ensure confidentiality, integrity, and availability of the information processed, stored, and transmitted by AI systems
10. AI solutions must include multidisciplinary stakeholder involvement or Interagency Coordination
   - All sectors affected by the AI solution should coordinate and share experiences and challenges of AI solutions

Safety, transparency and multidisciplinary aspects are key ingredients to a successful and well-thought out AI solution.

CITATIONS AND ADDITIONAL READINGS

FDA, U. 2019. “Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD).” FDA.

US Department of Health and Human Services. Software as a medical device (SAMD): Clinical evaluation. Guidance for industry and Food and Drug Administration Staff. 2017.

BEST ETHICAL PRACTICES

BEST ETHICAL PRACTICES – PROBLEM FORMULATION

In this section, we will learn about specific actions that have been recommended as best ethical practices in the development and deployment of machine learning-based AI in health care applications.

Our thinking about best ethical practices will start at the stage of **problem formulation** and determining the purpose of the AI system that you want to develop.

First, is the intent ethical, is its purpose to enhance the health and well-being of patients? Even if the intended purpose seems ethical, could there be negative consequences if misused? Increasingly we are seeing data scientists challenging the purpose and the uses of the products they are developing, whether the purpose is to target advertising to individuals using social media, or to identify individuals through facial recognition algorithms. For example, the Association of Computing Machinery, one of the leading professional organizations of computer scientists, has urged a suspension of the use of facial recognition technologies because of their potential for prejudicial impact on human and legal rights. The Association has also asserted that developers and operators, as well as users of facial recognition technology are accountable for these systems' use and misuse.

While these might seem like questions that don't need to be asked in the health care AI context because the answers are obvious, they are especially important to ask for AI developed by teams that have conflicting or competing interests. Health care settings are rife with such competing interests because they are operating under resource constraints, including constraints on finances, personnel, equipment and supplies, not to mention financial incentives to improve care. Together, these interests and incentives all create pressure to avoid certain kinds of patients, or avoid providing certain kinds of treatments. AI developers have to be vigilant about mitigating conflicts of interest. But how? We'll talk about this more specifically, but first, let's consider the issue of problem formulation.

Formulating the AI problem involves translating a high-level goal such as "improving patient care" into actionable questions that can be answered by available data. The challenge is that actionable questions and available data are usually limited, so a lot can get lost in translation, and practical constraints on problem formulation can lead to undetected errors and bias. A fairly common and seemingly straightforward task such as risk stratification is often fraught because risk can be defined in many different ways, and usually in ways that are not measured directly but through proxy variables. We have seen that the use of health care costs as a proxy for risk or health care needs has led to racial bias because risk scores generated from predictive models based on costs underestimated

the health care needs of Blacks as compared to Whites with the same risk score. Fortunately, such bias can be tested for, at least for variables for which you suspect underlying bias exists such as race or gender. In this example you would be looking at whether the relationship between risk scores generated by the model and the variable of interest, health needs, differed by race.

Similarly, risk stratification on the basis of cost as a proxy for health needs tends to be biased towards older people with complex chronic conditions at the end of life, when most health care costs are incurred, and against children with acute, potentially fatal but treatable conditions. If the question to be answered is "who needs the most care" it is important to remember that this question is not merely quantitative but also a values question that demands nuance. How is "care" defined? Does the question distinguish between acute and chronic care? How is "need" defined? If patients have untreatable conditions and therefore typically do not receive costly care, does that mean that a model should assign them low risk scores, or that they do not need care?

Questions to ask in problem formulation

- Is the purpose of the AI system to enhance the health and well-being of patients?
- Could there be unintended consequences if misused?
- What is the intent of the system – to identify, classify, predict?
- What is being identified, classified or predicted?
    - E.g. risk, disease, prognosis, costs, health care utilization, admission or readmission, treatment efficacy, decompensation
- How are these terms defined?

Closely related to problem formulation is the choice of **data**. Of course, data that are already available are the easiest to use. But that doesn't mean that these are the right data. For example, electronic health records might be plentiful but they lack a lot of information that could be important to predictive models, such as environmental exposures, diet, or socio- economic factors, all of which we know are highly determinative of health and disease. EHRs from one hospital or health system, or insurance claims data also often do not provide a longitudinal timeline of data points for a patient over time because patients often move between health providers and insurers. And relying on data from a single time point could be misleading. On the other hand, grouping data together can mask important patterns. In our example of risk stratification, a model trained on data from people of a mix of age ranges could be masking patterns in data from a subpopulation, such as younger people. Of course, detailed knowledge of differences between subpopulations is necessary to know whether customized models are required. How does one know this in advance?

Questions to ask about data

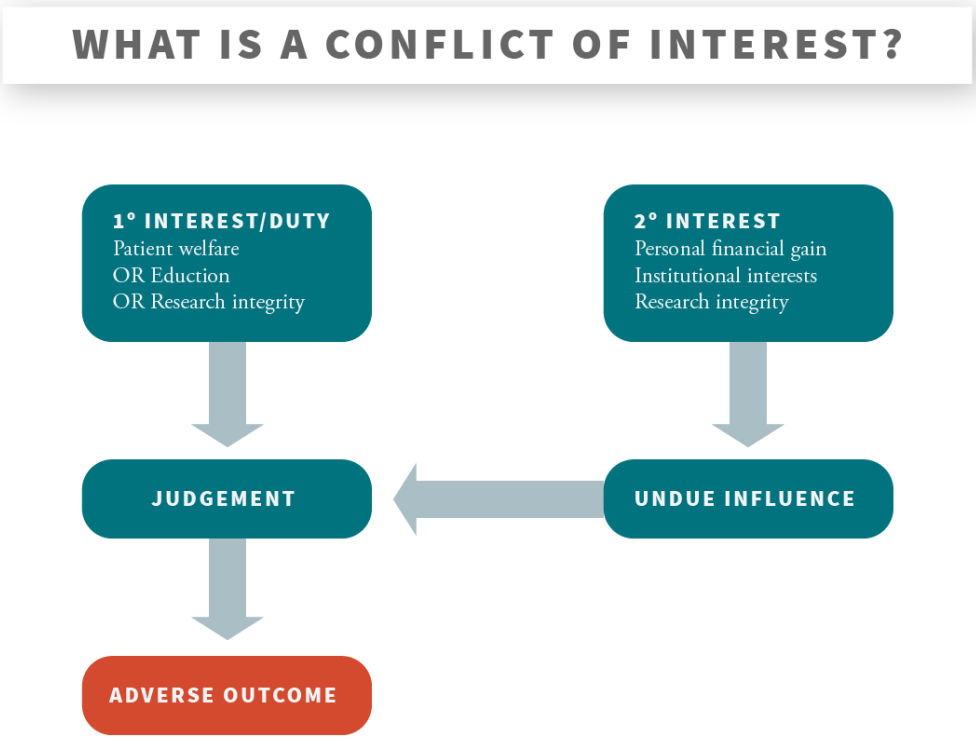
- How well are the variables in the model represented by available data?
- Are proxy measures being used?
- Are there important variables that are likely to be associated with main outcome measures that are not represented in available data?
- Are there likely to be differences between subpopulations in main outcomes, especially by legally protected characteristics such as age, race, ethnicity, or gender, or socially important characteristics such as income and education?

Best practices point to the need to include practicing clinicians who are intimately familiar with the relevant patient populations and conditions, and the data that are necessary for modeling They need to know what data are actually available, and the limitations of the data, especially in terms of likely biases. It is important to have team members with an understanding of the source of systematic error, such as whether error is introduced because of lack of data from specific subpopulations, physician biases (such as the lower likelihood of women with heart disease to get a diagnosis of heart disease as compared to men), or broader social inequities such as differential access to care. That is because understanding the source of systematic error indicates whether and how models can account for bias. Clinician team members can be critical at the problem formulation stage in particular, because deep clinical knowledge is so important to understanding the implications of asking questions in a particular way.

BEST ETHICAL PRACTICES – IDENTIFYING CONFLICTS OF INTEREST

In this section, I want to talk about best ethical practices that address the issue of conflicts of interest. In medicine and biomedical research, we deal with conflicts of interest all the time, and have developed ways of mitigating their negative effects. So, how is that done, and how can we apply those strategies to the development and deployment of AI in health care settings?

First, what is a **conflict of interest**? For our purposes, they exist only when there is a primary interest that is a duty. An example of a primary interest is the clinician's or hospital's duty to care for their patients. However, we all have multiple interests, some of which can compete or conflict with these interests. These other interests are called secondary interests, and can include, for example personal or institutional financial interests, duties to others such as people who are not the clinician's or hospital's own patients, or reputational interests, either positive or negative.



So, why are these secondary interests a problem? In the health care setting, the problem arises because in the course of carrying out duties, many decisions have to be made on behalf of patients, and decisions are often based on individual judgement. And judgement can be influenced by all of these secondary interests, in such a way as to subvert from the primary duty and cause harm to patients.

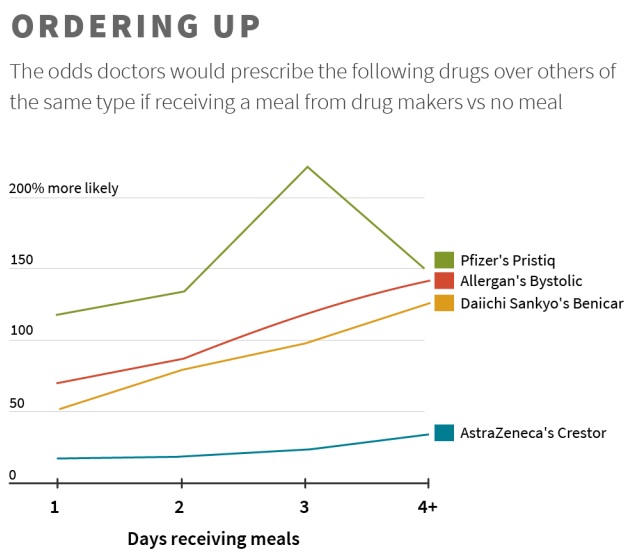
Photo by Pinar Kucuk on Unsplash

An example of this in clinical care is the influence of pharmaceutical sales representatives who try to persuade doctors to use their drugs instead of the competitor's drugs, when in some cases the competitor's drug might be better for the patient. Influence might be exerted in the form of financial incentives or gifts. These secondary interests have been linked to significant changes to prescribing practices, so they have real effects.

Although physicians think it is impossible to be influenced by accepting a slice of pizza paid for by a pharmaceutical company, there is a study demonstrating that prescriptions can be more than doubled specifically for drugs sold by the providers of meals over other similar drugs, with increases in prescribing seen even after just one

meal and rising with every additional day of meals provided. So, part of the problem is that these secondary interests have effects without our being aware of them.

In biomedical research, secondary interests can also include financial factors such as the promise of obtaining research grants or consulting fees from sponsors, or stock or royalties from companies whose products are being studied. Secondary interests can also be reputational or ideological, for example, a strong desire to gain fame or promote a specific hypothesis. In the case of research, the primary interest is usually the integrity of the research, although for research involving human subjects, the health and welfare of individual research participants is also of high priority. But what are the possible negative effects of influences of secondary interests?



Source: DeJong C, Aguilar T, Tseng C, Lin GA, Boscardin WJ, Dudley RA. Pharmaceutical Industry–Sponsored Meals and Physician Prescribing Patterns for Medicare Beneficiaries. *JAMA Intern Med.* 2016;176(8)

Again, let's look at all the **decisions** that are made in the course of research that could be subverted away from serving the primary interest of the integrity of the research. One is the **formulation of the research question**. Is the question in service of corporate interests or meeting a real patient need? For example, is a clinical trial designed to test a drug in order to serve patients who have no other therapeutic options, or a reformulation of an existing drug that will extend patent protection and a company's market position?

Another way that secondary influences can have effects is through creating unconscious bias, leading to inaccurate **measures of treatment effects**. We know that, in general, systematic error tends to inflate effect sizes. Clinical researchers have long recognized this, and therefore use techniques such as randomization and blinding to reduce bias, or systematic error. Choosing to use such techniques represent design decisions. Other important design decisions that could be influenced by secondary interests are what populations and data you choose to study, and **how you choose to analyze, report or share data and findings**.

Ways to Mitigate Conflicts of Interest:

Disclosure	• Publicizing the secondary interest
Mediation	• Putting financial (secondary) interest in blind trust • Putting some decisions about the primary interest in the hands of an independent body
Recusal	• Removing financial (secondary) interest or replacing the person with the primary interest

So, what do we do to mitigate potential effects of conflicts of interest? We can focus mitigation strategies on protecting the primary interest, eliminating or reducing the secondary interest, or both. The most common strategy, which you have probably employed or been asked to employ, is disclosure, which entails making public your secondary interests, such as disclosing speaking fees or research sponsorships from companies whose products are related to the primary interests. Although it is the most common strategy, it is also the weakest because it does nothing to reduce or remove the secondary interest, but relies on those to whom the interest is disclosed to understand the implications of the disclosure and act in some way to counteract possible negative effects. Other strategies that are more robust involve mediation of the primary or secondary interest, such as placing a financial interest in a blind trust, or having an independent oversight committee such as a Data Safety Monitoring Board in place to oversee or make important decisions, essentially taking them out of the control of the person who has the conflict of interests. The strongest strategies are recusal from secondary interests, such as selling one's stock, or even recusal from the primary interests, such as replacing an investigator with another person who does not have conflicting interests.

BEST ETHICAL PRACTICES – MITIGATING CONFLICTS OF INTEREST

In this section, we'll try to answer the question: how do we translate conflict of interest mitigation strategies to the development and deployment of AI for health care? Let's look at each of the three general types of strategies: disclosure, mediation, and recusal.

The principle underlying **disclosure** is transparency, or facilitating awareness of people who are impacted by AI and who are owed duties of care that other interests exist that could influence this care. For AI, this transparency is complicated by the fact that most of the people who could be impacted, such as patients and providers, are probably not even aware of the existence of AI or how it might be used in decision making about their health care. So first, some public notification of the use of AI may be warranted, especially if the application deviates substantially from commonly

understood or accepted uses of data in the health care context, or if the AI or data use poses more than usual risks or be especially sensitive. Examples might be using and storing video data from telehealth visits, or collecting and analyzing patients' social media data to predict or diagnose mental health conditions, or using EHR, sensor and claims data to guide decisions about offering palliative care at the end of life. This notification could also include disclosure of how AI is used in decision-making about patient care and who developed the AI system.

Other ways of implementing the principle of transparency include thorough reporting of how models were built, as suggested by members of the data science community. This includes clear descriptions of data sources, participants, outcomes and predictors, the contexts in which the model was validated, limitations and contraindications for deployment and assumptions or conditions that must be satisfied. The Association of Computing Machinery has also recommended that for facial recognition technologies, or other uses of AI where racial or other biases are of concern, that error rates be reported disaggregated by race, gender and other context-dependent demographic features.

The principle behind the strategy of **mediation** is independent review or oversight. For AI development, this could be achieved by auditing of algorithms and models by third parties. This process was suggested by the Obama administration in 2016 to mitigate discriminatory practices and civil rights violations. Employing algorithmic audits also enhances transparency to the extent that it encourages developers to make algorithms auditable in the first place.

The strategy of **recusal** in the context of AI development is difficult to implement. If AI development is being conducted in an academic research setting, it might be possible for the developers to recuse themselves of financial interests, or replace people with financial interests with others who are more independent. However, in a corporate setting, that might be impossible, so developers would have to rely on disclosure or mediation-based strategies, keeping in mind that disclosure-based strategies are very weak and do little to engender or maintain trust of key stakeholders.

The takeaway points here are that:

- Best ethical practices in developing AI for health care include careful formulation of the problem to be solved, ideally with input from people who have deep knowledge of the specific clinical settings and data relevant to the problem.
- Conflicting interests can have real impacts on design decisions, but there are strategies to mitigate the potential negative effects of conflicts of interests.