Stanford MEDICINE

Stanford | Center for Health Education

EVALUATIONS OF AI APPLICATIONS IN HEALTHCARE STUDY GUIDE

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

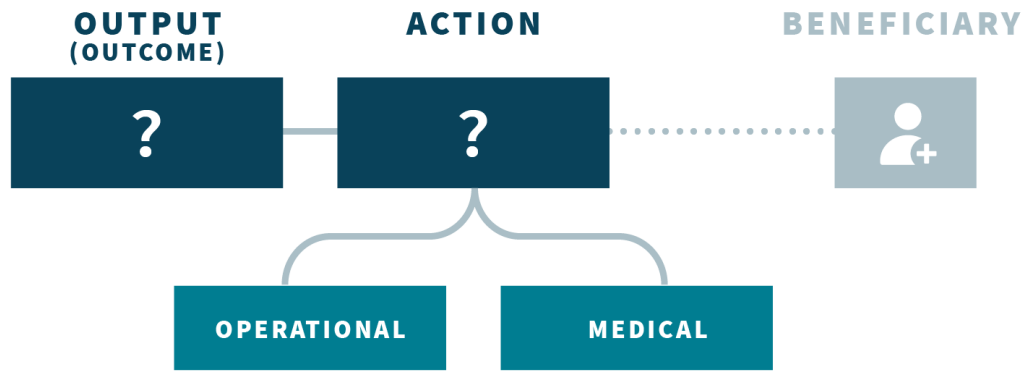
Considerations when evaluating an AI solution:

- 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.

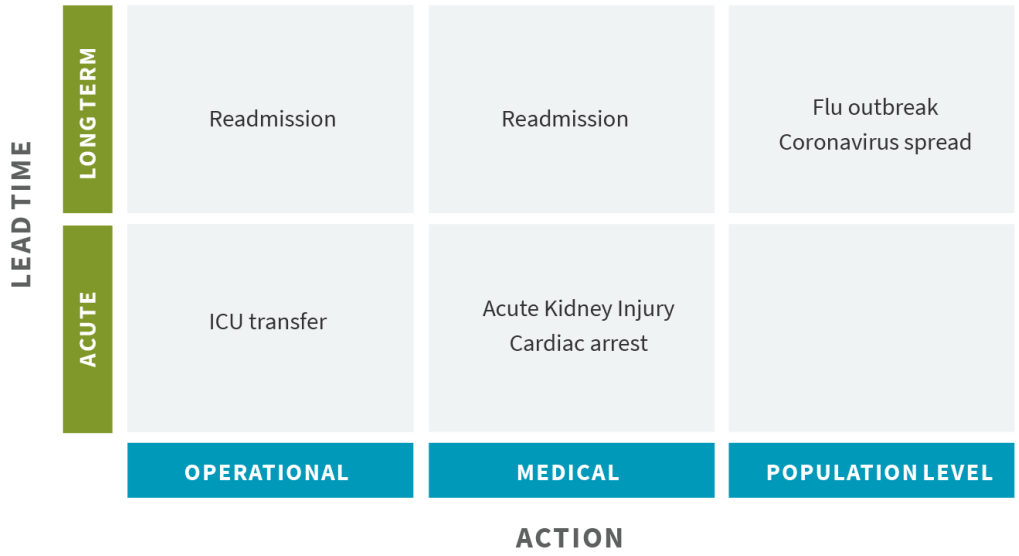
When evaluating an AI solution, it is also important to think about the lead time your action needs. One must think about whether the action to my output is acute (it has to happen immediately), or is it long-term?

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.

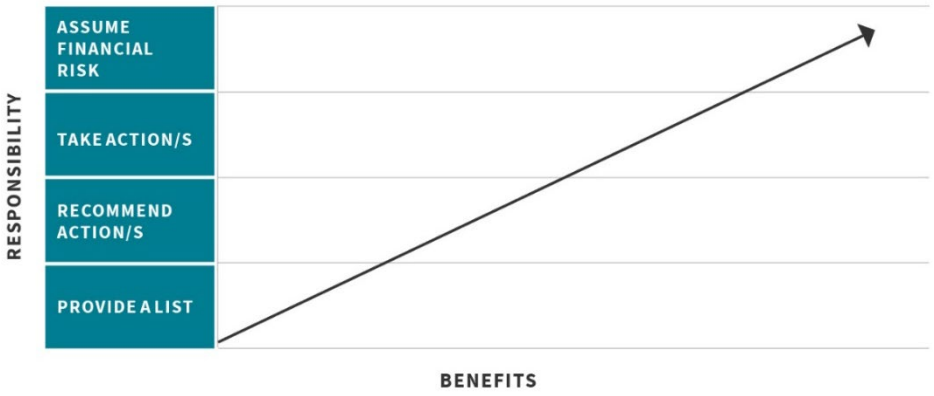
We need to understand if the action is **operational** or **medical** - and an AI solution can be developed to have either action.



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.

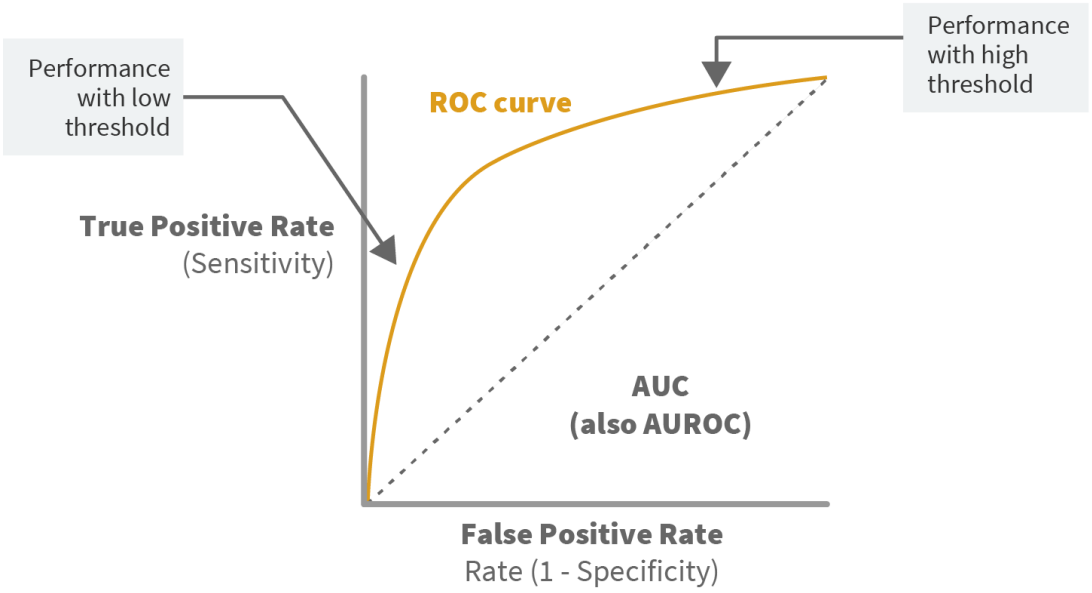
CLINICAL UTILITY

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.



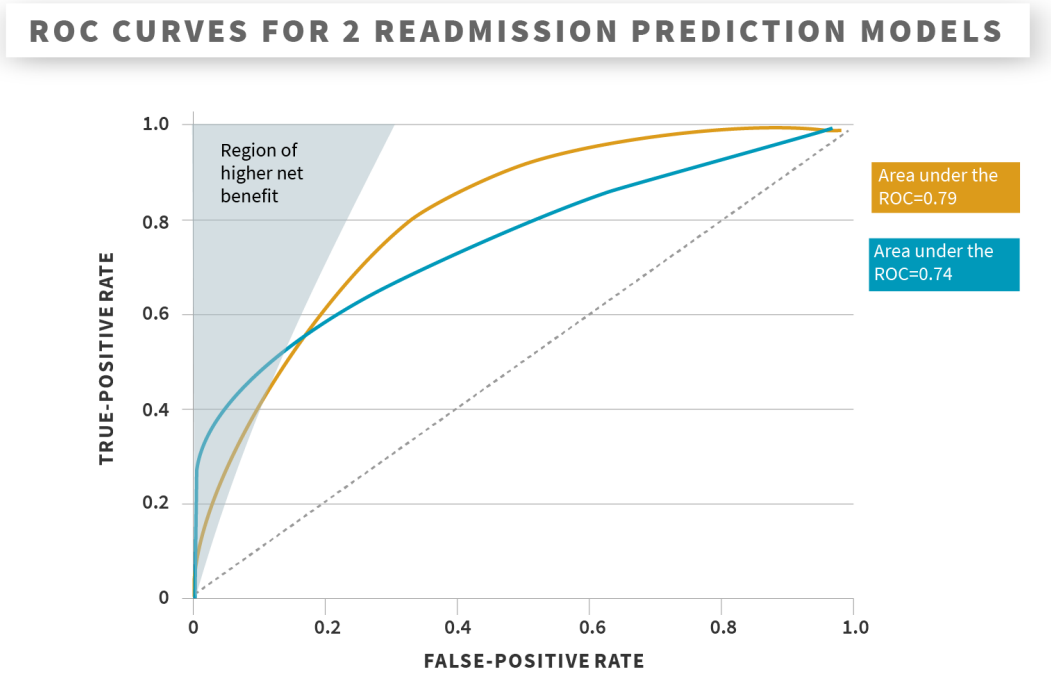
The ROC curve is a plot of the true positive rate against the false positive rate at different threshold settings. The true-positive rate is also known as sensitivity or recall. The false-positive rate is also known as probability of false alarm, which is 1 – specificity.

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