EVALUATIONS OF AI APPLICATIONS IN HEALTHCARE STUDY GUIDE

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

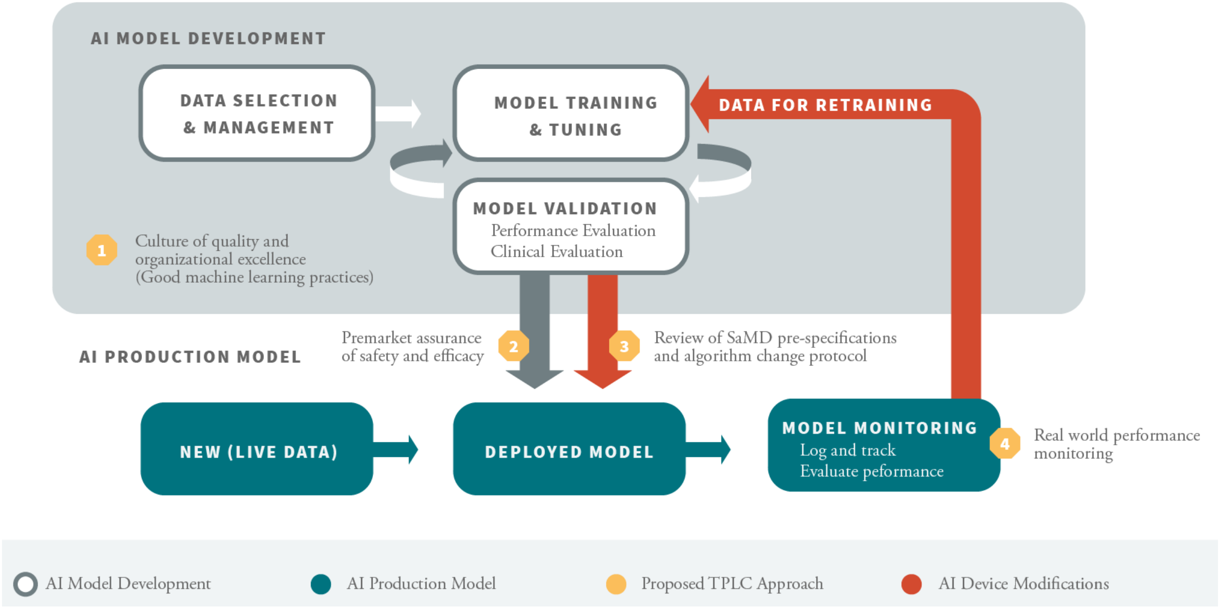
AI in healthcare has moved slower because human lives are at stake and therefore regulations, policies, and standards have set a higher bar for product development and deployment. Enthusiasm for AI in healthcare has been overshadowed by the challenging path to successfully implement these solutions into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

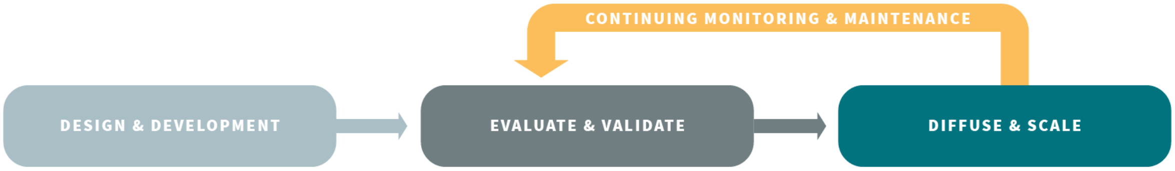
Once you have completed the first section “AI Model Development”, you move to the AI Production Model, or the “Deployed Model” which continuously receives new (or live) data and is continuously monitored for safety and performance evaluations.

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.