Stanford MEDICINE

Stanford | Center for Health Education

FUNDAMENTALS OF MACHINE LEARNING FOR HEALTHCARE

MODULE 6 - BEST PRACTICES, TEAMS, AND LAUNCHING YOUR MACHINE LEARNING JOURNEY

LEARNING OBJECTIVES

- Describe best practices for developing and evaluating clinical machine learning applications
- Understand the "output-action" pairing as a framework related to machine learning in healthcare applications
- Learn what skillsets are useful for multidisciplinary and diverse teams and how each contributes to success
- Recognize the basic challenges around regulatory and ethics related in clinical machine learning
- Become familiar with challenges with human computer interaction for machine learning applications including automation bias and the consequences in healthcare
- Analyze the potential impact on the future clinical workforce by machine learning on delivery of healthcare

DESIGNING AND EVALUATING CLINICAL MACHINE LEARNING APPLICATIONS

Understanding of clinical machine learning project development - The name of the game is to "find problems worth solving".

One of the ways to understand the value of the potential solution is to consider all actions and repercussions that would come from a solution - and this is where having input from multiple stakeholders and domain experts is important.

FINDING PROBLEMS WORTH SOLVING

	SCIENCE	PRACTICE	DELIVERY
CLASSIFY	Finding sybtypes of heart failure with preserved injection fraction	Who might be at high risk for a thromboembolism?	Who is burnt out?
PREDICT	Estimating the disease risk conferred by genetic variations	Which patients are at risk of dying in the next 3-12 months?	Who will be a no show?
ACT/TREAT	XYZ solid tumors can be treated by allogeneic chimeric antigen receptor T-cell By	What is a good second line drug to manage diabetes after metformin?	Request four back up nurses on Wed, for the Ortho OR.

Three categories of model application:

- Scientific exploration / discovery
- Clinical care / decision support
- Care delivery / managing medical practice

Three categories of model output we think about:

- Classification, prediction, recommendation

The Output Action Pair (OAP): The action that will result from the model output.

- Consider what a correct model prediction will entail
- Consider what an incorrect model prediction will entail

Without the right problem, it doesn't matter if you have the best scientists, infinite computational resources, or the most perfect dataset.

Utilizing the OAP Framework:

1. Suppose we would like to design a machine learning project to reduce deaths from sepsis in the ICU

2. Assume we have a prediction + practice sepsis model that can ingest real-time patient EMR data, and then predict the patients likely to develop sepsis at the point of care.
    - The output will be the prediction and the action will be an alert to the clinical team if positive above a set threshold.
3. Output here will be a sepsis diagnosis, which will be the label that we will use to build our model. What is the definition of sepsis?
    - In this case sepsis has distinctly different definitions depending on what you are attempting to address.
    - One definition is the Sepsis-3: This is a medically accurate consensus definition that uses specific clinical criteria and applied to a patient as a formal diagnosis
    - Another definition of sepsis is the Medicare sepsis identifier SEP-1.
        - A quality measure used by Medicare and Insurance reporting for billing and quality reporting
        - In contrast to the other definition, this measure does not represent medically useful sepsis definitions
        - This definition considers only a subsample of patients in a given hospital so you would not have all the sepsis patients labeled with this approach
4. Be sure that the labeling procedure matches the objective of your model

OUTPUT	ACTION/COST/RESULT
TRUE	Escalation in sepsis related medical therapy - possible improved outcome - increased costs of care possibly outweighed by earlier intervention - avoid costs/consequences of untreated sepsis
TRUE NEGATIVE	Avoid unneeded sepsis related medical therapy - no expected change in outcome - reduce unneeded costs
FALSE POSITIVE	Escalation in medical therapy - unclear impact on clinical outcome - increased costs of care - detrimental impact on model effectiveness/adoption
FALSE NEGATIVE	Avoid needed sepsis related medical therapy - worsened outcome related to delayed sepsis therapy - increased cost of untreated sepsis - detrimental impact on model effectiveness/adoption

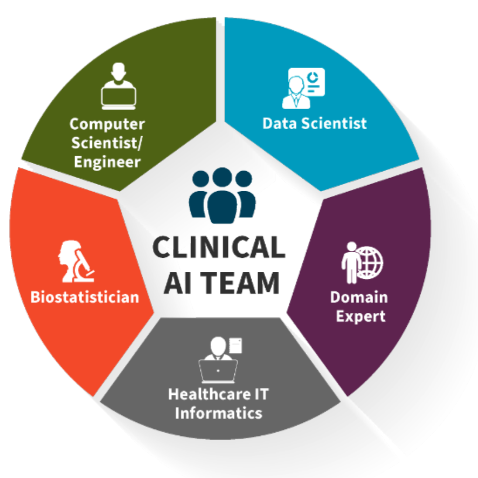
5. OAP utility analysis affords a rough understanding of the minimum acceptable performance and how the output would lead to action in many possible scenarios

6. Consider how humans will interact with the model in production
    - Models are often evaluated using statistical metrics for success:
        - Positive predictive value, sensitivity/recall, specificity, calibration
    - Factors to consider:
        - Lead time offered by the prediction
        - The existence of a mitigating action
        - The cost intervening and the cost of false positives and negatives
        - The logistics of the intervention
        - Incentives both for the individual and the healthcare system
        - Alert fatigue
        - Cognitive biases
    - Models could lead to complacency from those who begin to trust the model too strongly

It is important to build a team with diverse expertise:

- Build a team with expertise in: clinical medicine, clinical trials, statistical study design, healthcare finance and incentives, data primary and biases, end user environment
- Not everyone on the team needs to be able to write out the math that explains how weights are updated in backpropagation
- The entire team benefits if everyone has a high-level understanding of machine learning concepts and principles because that foundation of knowledge serves as a common language allowing everyone's unique expertise and experience to apply to the problem

Archetype areas of expertise:



- Data scientist
  - Focused on data mining, feature engineering, analytics, metrics of model(s) performance
  - A deep knowledge working with healthcare data in this role is critically important because this role requires delivering and manipulating the data that the rest of the team can work with
  - Feature engineering, pre-processing, and other tasks that will be key to a successful project like preliminary simple model building to get a sense of which machine learning approach is best or what features to use or not use
- Machine learning engineer
  - An expert in computer science, that would ideally team up alongside the data scientist co-developing the model
  - Focusing on the machine learning techniques needed to obtain high-performing models
  - May also play a leading role in writing the more formal code for final software deployment, and the entire workflow pipeline, in other words building out formal, production-ready models and setting up the tools to integrate them with the rest of the clinical enterprise
  - Often the machine learning engineer is knowledgeable in more advanced machine learning techniques, especially deep learning, computer vision, and natural language processing
- Statistician
  - Help form conclusions safely beyond the data analyzed and back that up with either trial design or statistical analyses
  - In a lot of circumstances, the statistical skills sound a lot like our discussion of data scientists evaluating pilot model performance - after all there is a lot of statistical knowledge needed to evaluate models and review metrics
- Healthcare IT
  - Integration and deployment in a healthcare environment. It is incredibly common for teams to build models that work only to hit a long delay in integration
  - It is important especially for models that are geared toward clinical deployment to engage healthcare IT professionals early in the model development process
    - Knowledgeable about the details of when and where certain data become available, whether the mechanics of data availability and access are compatible with the model being constructed, and the important interactions within the existing healthcare ecosystem

- Domain expert
    - Provide context and guide the development of the overall application, help decide metrics, and make key development decisions including where to choose a threshold
    - What populations and data should be included, and how deployment might take shape?

Domain Experts	Category	Examples of Applications
Device product developers, clinicians, end users (patients and families)	Health monitoring Benefit/risk assessment	Devices and wearables Smartphone and tablet apps, websites
	Disease prevention and management	Obesity reduction Diabetes prevention and management Emotional and mental health support
	Medication management	Medication adherence
	Rehabilitation	Stroke rehabilitation using apps and robots
Clinician care teams	Early detection, prediction, and diagnostics tools	Imaging for cardiac arrhythmia detection, retinopathy Early cancer detection (e.g., melanoma)
	Surgical procedures	Remote-controlled robotic surgery AI-supported surgical roadmaps
	Precision medicine	Personalized chemotherapy treatment
	Patient safety	Early detection of sepsis
Public health program managers	Identification of individuals at risk	Suicide risk identification using social media

	Population health	Eldercare monitoring
	Population health	Pollution epidemiology Water microbe detection
Healthcare administrators	International Classification of Diseases, 10th Rev. (ICD-10) coding	Automatic coding of medical records for reimbursement
Healthcare administrators	Fraud detection	Health care billing fraud Detection of illegal prescribing patterns
Healthcare administrators	Cybersecurity	Protection of personal health information
Healthcare administrators	Physician management	Assessment of quality of care, outcomes, billing
Geneticist	Genomics	Analysis of tumor genomics
Pharmacologist	Drug Discovery	Drug discovery and design

GOVERNANCE, ETHICS, AND BEST PRACTICES

At the very least there should be a plan for medical data stewardship that everyone involved in the project can agree to. It is critical that all members of the team be trained and follow strict best practices when working with healthcare data, even if de-identified, since a breach or leakage of data could be catastrophic to the project.

Medical data stewardship can be very different than other forms of data.

Up front training for all who are involved can ensure that ensure that everyone involved in the project has had at least basic medical research and data stewardship training.

- Be careful of the 'context transgressions' that can occur in collaborations or partnership where data may flow between medical, social and commercial contexts governed by different privacy norms

De-identified patient data is not considered private health information because of the anonymization process. However, research has shown that it may be possible to re-identify de-identified patients given the right kind of data.

When curating large medical datasets for clinical machine learning applications, it is important to ensure that the data will not be used in a way that could cause harm or is unethical. Members of the development team and ecosystem must share a common understanding of the ethical and regulatory issues around the development and use of clinical machine learning tools.

When selecting members of the team it is critically important to consider how the project will promote equitable clinical machine learning solutions that does not suffer from bias.

- Working to ensure that new clinical machine learning applications are free from biases can include the composition of the team, and forming a team that is diverse with respect to gender, culture, race, age, ability, ethnicity, sexual orientation, socioeconomic status, privilege, etc.

The development of health care AI tools requires a diverse team with varying skillsets:

- Information technologists
- Data scientists
- Ethicists and lawyers
- Clinicians
- Patients
- Clinical teams
- Organizations

These teams will need a macro understanding of the data flows, transformations, incentives, levers, and frameworks for algorithm development and validation, as well as knowledge of ongoing changes required post-implementation.

Consider machine learning tools that, if successful, can achieve significant cost reductions or improved patient outcomes and create further competitive advantage and exacerbate existing healthcare disparities.

Best practices should be driven by an implementation science research agenda and should engage stakeholders with an effort to reduce cost and complexity of AI technologies. In summary - the best clinical machine learning outcomes will come from team-based approaches composed of individuals with complimentary skillsets, essential expertise, and diversity of backgrounds.

HUMAN FACTORS IN CLINICAL MACHINE LEARNING - FROM JOB DISPLACEMENT TO AUTOMATION BIAS

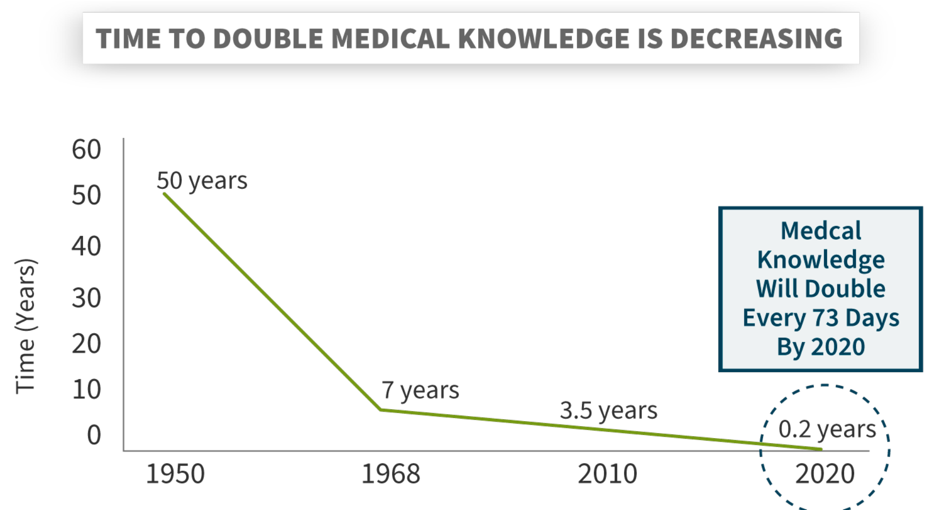
Tasks that can be automated:

- Transcribing clinical documentation
- Image analysis
- Billing and coding
- Practice management
- Staffing and resources optimization
- Prior authorization forms
- Triaging routine diagnosis

Machine learning applications remain unlikely to displace many human skills such as complex reasoning, judgment, analogy-based learning, abstract problem solving, physical interactions, empathy, communication, counseling, and implicit observation.

- Gradually there will thus likely be a shift in health care toward jobs that require direct physical(human) interaction, including specialized surgical or procedural specialties, which are not so easily automated

It is increasingly evident that a transition into the machine learning era of health care will require education and redesign of training programs.



Graphic source, NCBI, “Challenges and opportunities facing medical eduction”
Peter Densen, MD,2011

- Medical training institutions already acknowledge that emphasizing rote memorization and repetition of information is suboptimal
- Medical knowledge doubles every 3 months on average

Jobs could be lost or replaced, but the process of the "machine learning healthcare" era will create new jobs for scientists and engineers who understand healthcare and healthcare professionals who understand machine learning.

Important aspects of this new machine learning medical hybrid profession will include:

- Training machine learning systems and deliberate effort to evaluate and stress test them
- Leading multi-disciplinary teams within the healthcare system to provide ongoing machine learning education and guide strategy for clinical practice
- Ongoing evaluation and testing of machine learning models in the active healthcare environment, in particular due to the dynamic and changing healthcare landscape

Education for the healthcare workforce:

- Principles and impacts of machine learning
- How to interpret the model recommendations
- The flaws and biases
- How to identify unintended consequences of machine learning system behavior

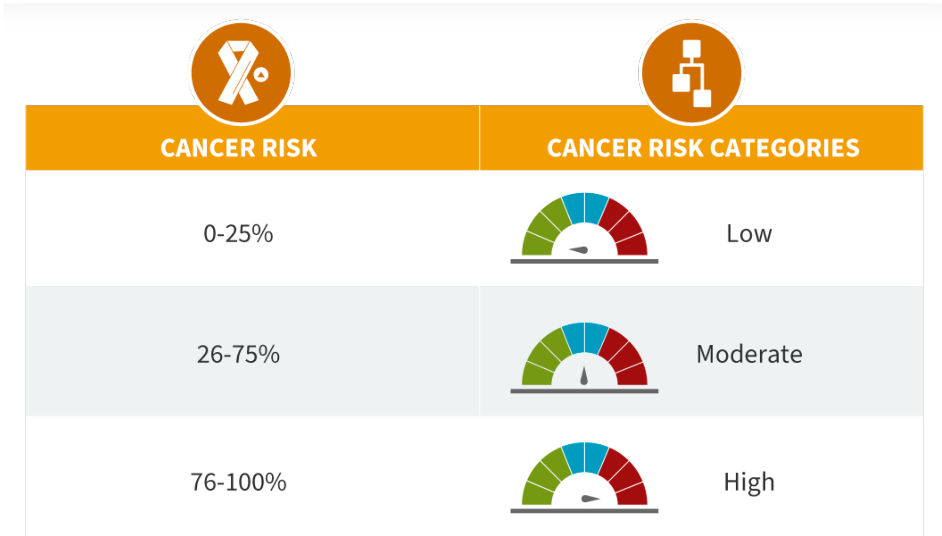
There are further risks of machine learning in healthcare which could negatively impact the field

- **Deskilling (aka "skill rot")**: A risk of over-reliance on computer-based systems for cognitive work

"Automation bias" refers to the fact that when humans have the guidance of automated systems they begin to act as if they are in a lower risk environment.

- **Automation misuse**: When a healthcare work's inherent trust in an automated system leads to overreliance on automated aids resulting in a decreased performance
- The airline industry has found that the design of many decision support systems has contributed to problems such as automation bias and automation misuse. The pilots performed best when the automated system recommendation was presented with a trend display of system confidence
- Engineering systems that could provide a recommendation AND the system probability calculation provides best performance

Calibrated risk could allow for better clinical decision making:



CANCER RISK	CANCER RISK CATEGORIES
0-25%	Low
26-75%	Moderate
76-100%	High

- More specific numerical risk scores may provide better information and lead to more nuanced treatment discussions than predicting simply "low risk"
- A calibrated continuous risk estimate is more informative and allows the user to set their own thresholds or decide if they will "trust" the output - avoiding the hazards of automation bias and misuse

CITATIONS AND ADDITIONAL READINGS

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, *17*(1), 195. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6821018/

Office, U. (2020, January 21). Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development [Reissued with revisions on Jan. 31, 2020.]. https://www.gao.gov/products/GAO-20-215SP

Wynants, L., van Smeden, M., McLernon, D.J. *et al.* Three myths about risk thresholds for prediction models. *BMC Med* **17**, 192 (2019). https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1425-3