It’s not just a last mile problem.

Making machine learning work in the real world of patient care

Ron Li, MD[1]
Naveen Muthu, MD[2]
Jonathan Chen, MD PhD[1]
Swaminathan Kandaswamy, PhD [3]
Margaret Smith, MBA[1]

[1]Stanford University School of Medicine,
[2]Children’s Hospital of Philadelphia (CHOP),
[3]Emory University School of Medicine.

Panel Members

Ron Li, MD
Clinical Assistant Professor of Medicine, Stanford School of Medicine

Medical Informatics Director for AI Clinical Integration, Stanford Health Care

Naveen Muthu, MD
Assistant Professor of Pediatrics, University of Pennsylvania Perelman School of Medicine

Human-Systems Informatics Lab Director, CHOP Department of Biomedical and Health Informatics

Swaminathan Kandaswamy, PhD
Instructor, Emory University. School of Medicine

Lead Human Factors Engineer, Cognitive Informatics Lab, Department of Pediatrics

Jonathan H. Chen, MD PhD
Assistant Professor of Medicine, Stanford School of Medicine

Center for Biomedical Informatics Research + Division of Hospital Medicine

Margaret Smith, MBA
Director of Operations, Stanford School of Medicine

Stanford Healthcare AI Applied Research Team, Department of Medicine

Disclosure

The members of this panel have no relevant relationships with commercial interests to disclose.

Learning Objectives

- Understand current barriers to improving healthcare using machine learning

- Key concepts in data science, process improvement, and human factors as they related to a machine learning implementation project

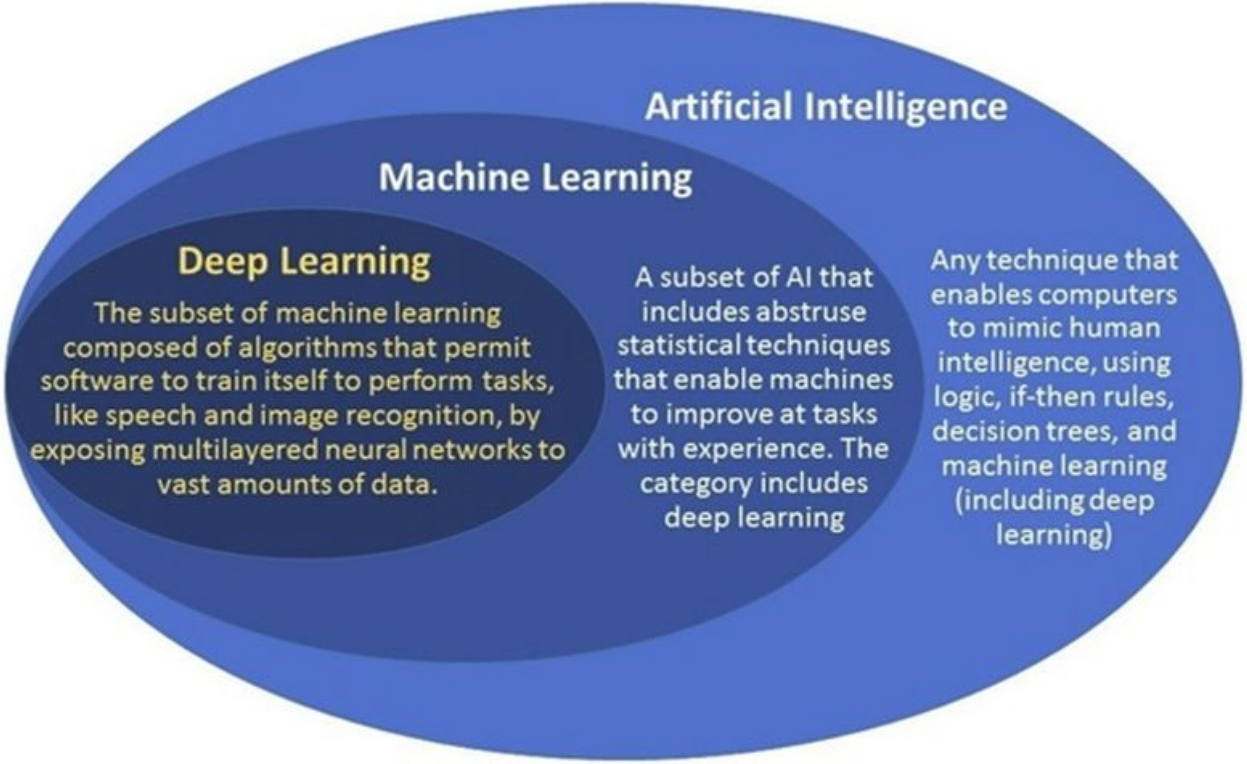
- How a multidisciplinary team can be leveraged to develop a machine learning solution for a complex healthcare problem

You are called by hospital leadership . . .

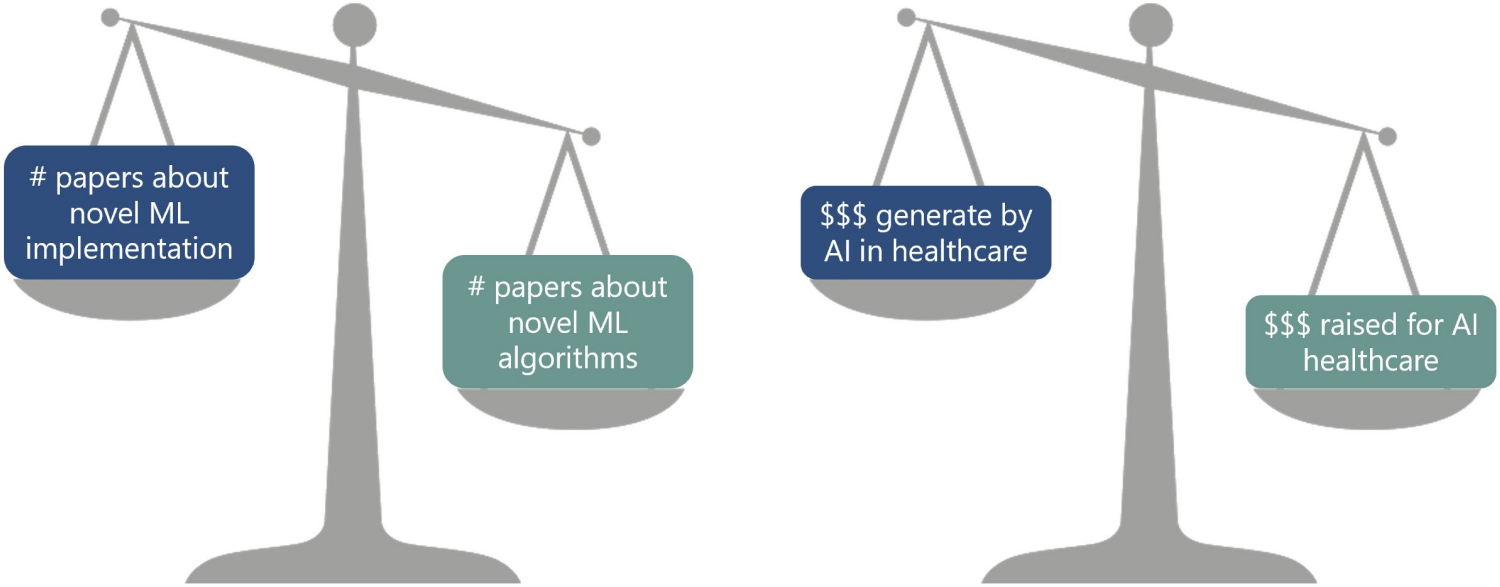
"We heard about some AI models out there that can predict death and ICU transfers. How do we turn it on in the EHR? We want to start using it at our hospital."

As the clinical informaticist leading your hospital's AI implementation team, how would you approach this request?

Artificial intelligence vs. machine learning



AI and ML have not translated into real world solutions for healthcare



Artificial intelligence / Machine learning

Google's medical AI was super accurate in a lab. Real life was a different story.

If AI is really going to make a difference to patients we need to know how it works when real humans get their hands on it, in real situations.

by **Will Douglas Heaven** April 27, 2020

"*Sounds* impressive. But an accuracy assessment from a lab goes only so far. It says nothing of **how the AI will perform in the chaos of a real-world environment.**"

https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/

How can we change our thinking . . .

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

From…

Ok, I have a machine learning model…now what?

To…

Ok, I have this problem I need to solve…how could machine learning enable the solution?

npj | Digital Medicine

www.nature.com/npjdigitalmed

COMMENT OPEN

Check for updates

Developing a delivery science for artificial intelligence in healthcare

Ron C. Li[1,2✉], Steven M. Asch[3,4] and Nigam H. Shah[2]

"To address how AI can be leveraged at scale, we need to both broaden and deepen our thinking around **how AI fits into the complexities of healthcare delivery**."

Li, R.C., Asch, S.M. & Shah, N.H. Developing a delivery science for artificial intelligence in healthcare. *npj Digit. Med.* **3,** 107 (2020).

Healthcare delivery occurs in a *complex* system

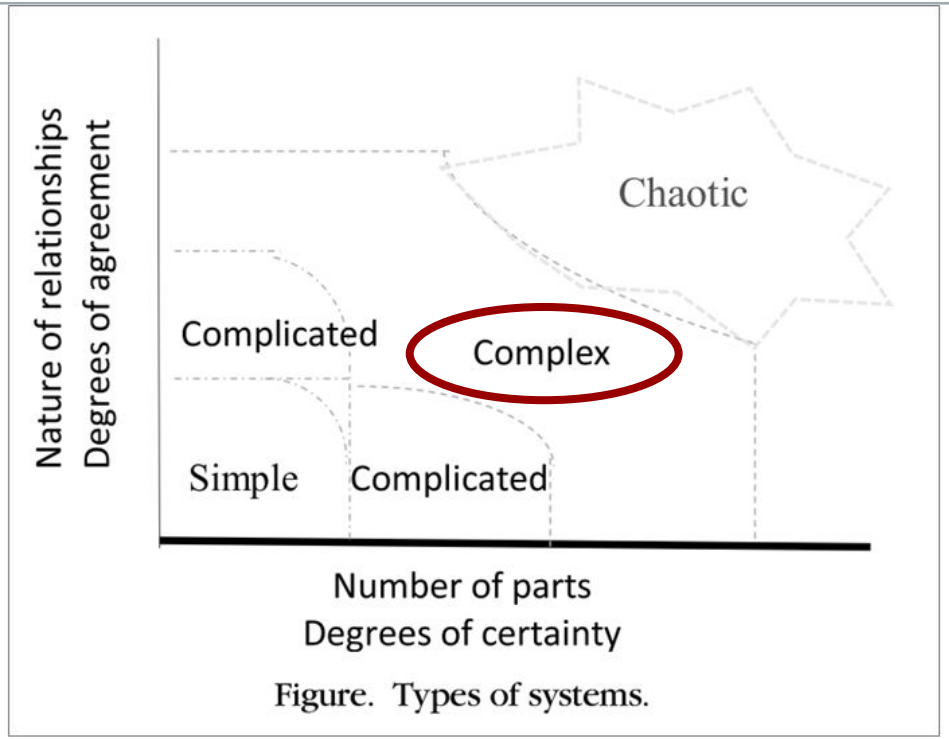
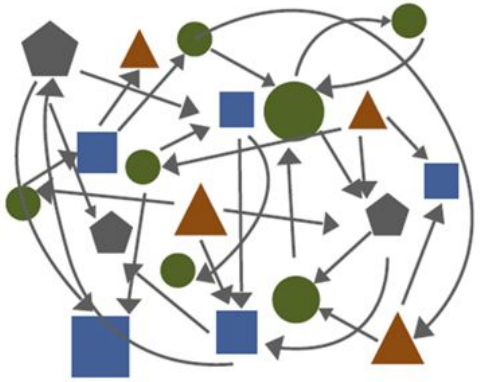


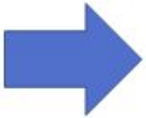
Figure. Types of systems.



System: a set of individual agents interacting with each other and forming **structures, processes, and patterns**

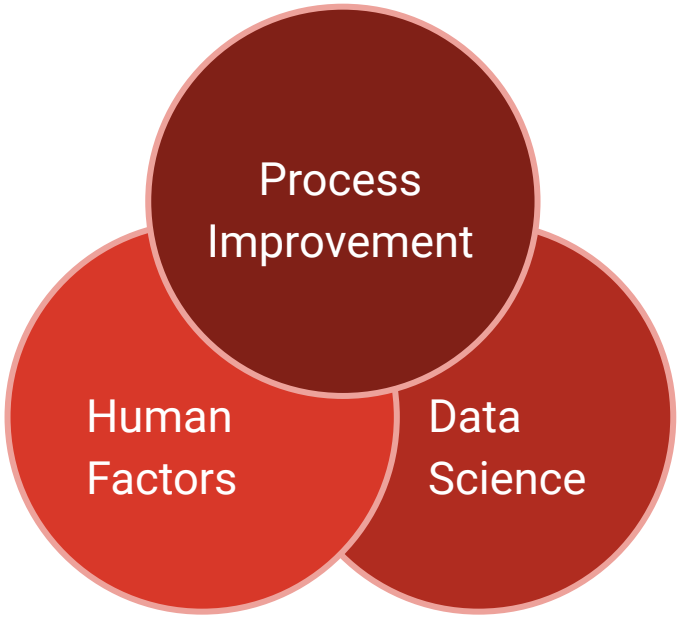
Ratnapalan S, Lang D. Health Care Organizations as Complex Adaptive Systems. *Health Care Manag (Frederick).* 2020;39(1):18-23

Creating systems enabled by machine learning . . .

Machine learning model(s)  Intelligent care delivery system

*(New work structures and processes for delivering care **enabled** by ML models)*

. . . requires the synthesis of multiple disciplines

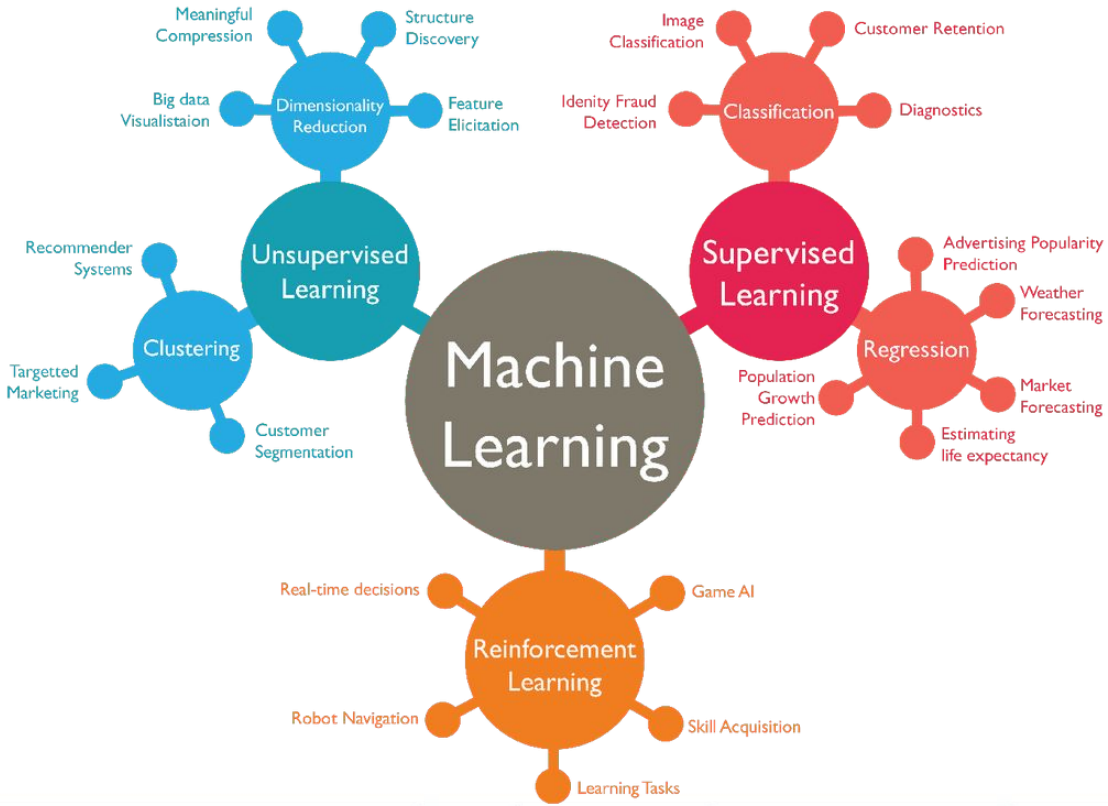


Implementing Machine Learning in Medicine
Data Science Perspective

Actionable	Arbitrary	Ascertainable
Viable	Variable	Verifiable
Important	Human	Correct
Decision	Prediction	Result

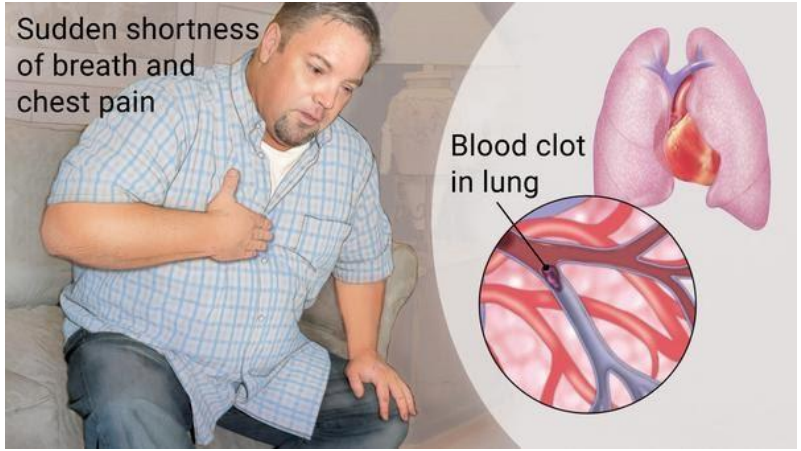
Jonathan H. Chen MD, PhD
@jonc101x
Center for Biomedical Informatics Research
+ Division of Hospital Medicine
Department of Medicine, Stanford University

Machine Learning Breakdown

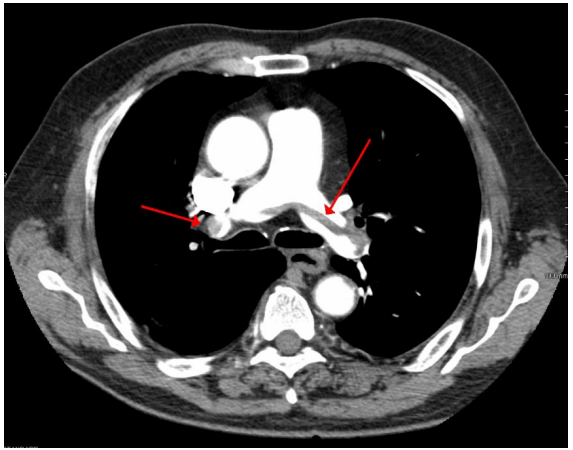


http://www.isaziconsulting.co.za/machinelearning.html

Prediction / Prognosis Example



Google.com

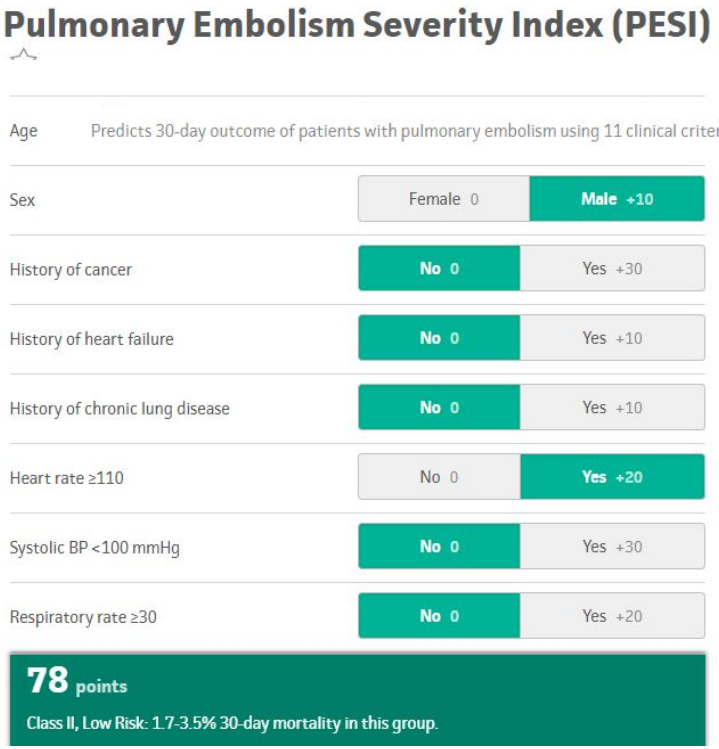


James Heilman, MD

Is this patient going to be okay?

Should I admit him to the hospital?

Prediction Examples



Pulmonary Embolism Severity Index (PESI)

Age — Predicts 30-day outcome of patients with pulmonary embolism using 11 clinical criteria.

Sex	Female 0	**Male +10**
History of cancer	**No 0**	Yes +30
History of heart failure	**No 0**	Yes +10
History of chronic lung disease	**No 0**	Yes +10
Heart rate ≥110	No 0	**Yes +20**
Systolic BP <100 mmHg	**No 0**	Yes +30
Respiratory rate ≥30	**No 0**	Yes +20

78 points

Class II, Low Risk: 1.7-3.5% 30-day mortality in this group.

Risk Scores (Manual):

- CHADS^2, ASCVD,
- Wells', APACHE

Non-Medical:

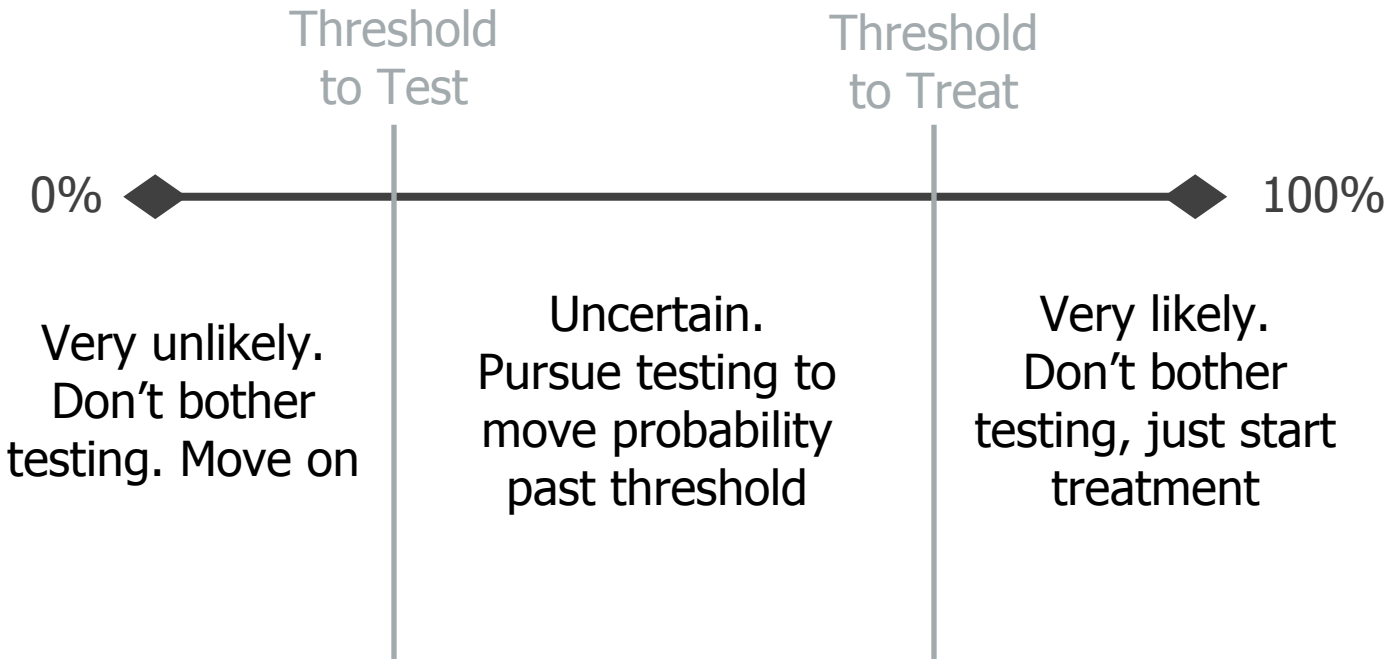
- FICO Score
- Bail Determination
- Spam Filter

Automated Scores:

- Sepsis Early Warning
- Mortality + Utilization
- Discharge Timing
- OR Scheduling

http://www.mdcalc.com

Threshold Approach Med Decisions



Predictive Model Opportunity?

Actionable	Arbitrary	Ascertainable
Viable	Variable	Verifiable
Important	Human	Correct
Decision	Prediction	Result

Predict death in palliative/terminal/ICU patients? – Actionable?
Administer oxygen for hypoxic patient? – Arbitrary?
Diversion of prescribed opioids? – Ascertainable?
Allocate a Scarce Resource or Risky Treatment

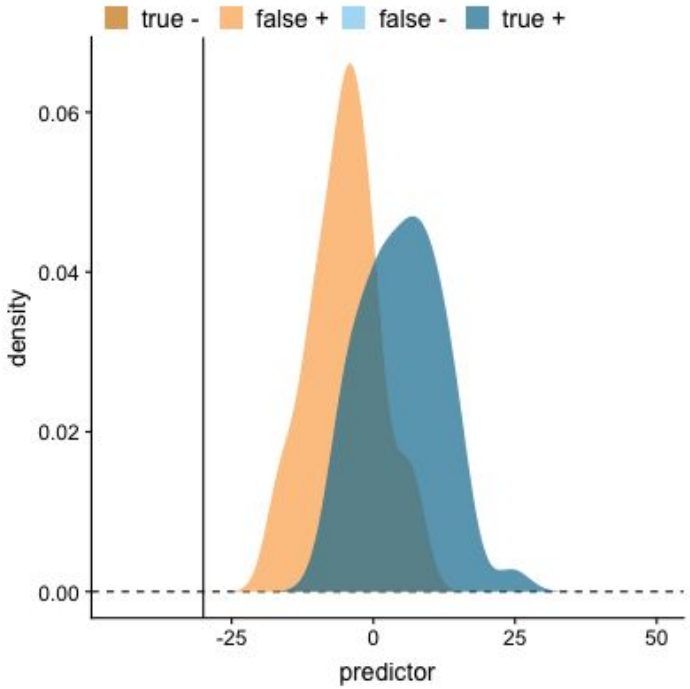
- E.g., Organ Transplant Prioritization or Prophylactic Medications

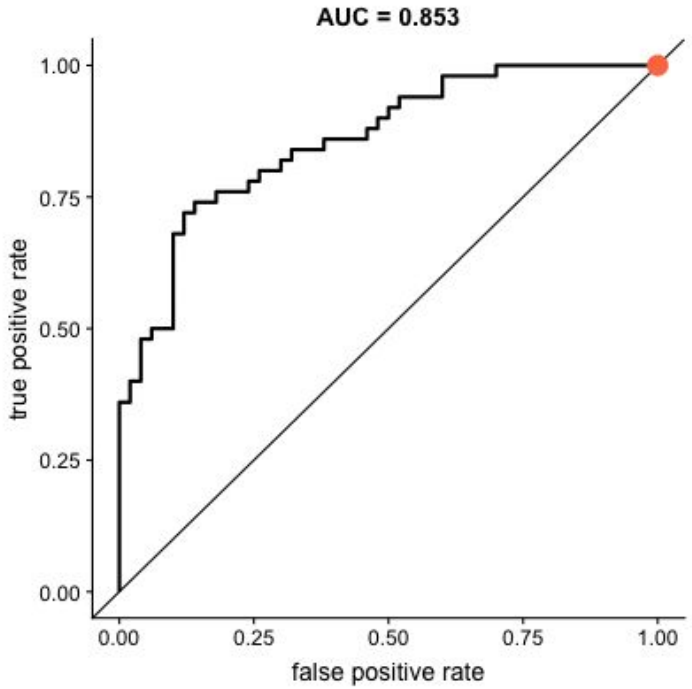
Prediction as Diagnostic Test Interpretation

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

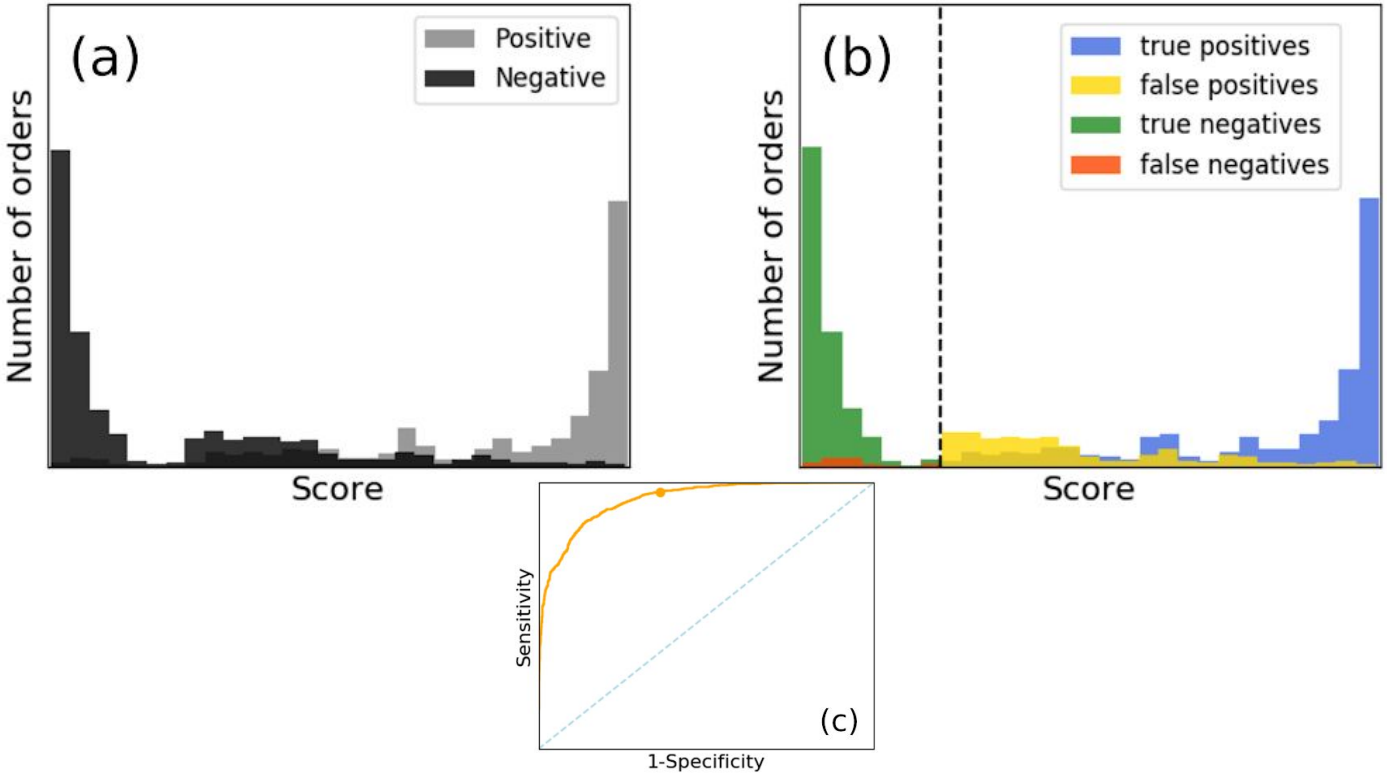
Visualizing Decision Thresholds





https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

Class + Risk Imbalance Decision Threshold



Problem Selection Principles

Do not want Technology, want Solutions to Problems

- Domain expertise, collaborations, workflows

Arbitrary? Decision depends on (Human) Estimation

- Separate prediction from action / intervention

Ascertainable? Data Source > Specific Algorithm

- Clear Outcome / Target / Feedback Available?

Actionability? Who can respond and how?

- Where to deploy an existing scarce or risky resource

Explanation essential or over-rated?

- Really need confidence / trust. Explanations are a crutch

Incentives are everything / Stakeholders?

- Value-based payment in flux. 90% still fee-for-service

Acknowledgements

- Students + Collaborators
- NIH / National Library of Medicine R56LM013365
- National Science Foundation #1914373
- Gordon and Betty Moore Foundation GBMF8040
- Stanford Clinical Excellence Research Center
- Stanford Departments of Medicine and Pathology
- STARR Clinical Data Warehouse
  - Stanford Medicine Research IT, School of Medicine Research Office

Content does not necessarily represent the views of the NIH, VA, or Stanford Health Care
Other financial disclosures: Reaction Explorer LLC – Co-founder
Consulting for NIDA Clinical Trials Network, Tuolc Inc., Roche Inc., Younker Hyde MacFarlane LLC

AI Implementation Using **Process Improvement Methods**

Margaret Smith, MBA

Director of Operations, Stanford Healthcare AI Applied Research Team

Outline

- What's the problem with jumping to solutions?
- What is Process Improvement?
- Relevant concepts/methods
- Why they are useful for AI implementations

Why do we jump to solutions?

"Don't bring me more problems! Bring me solutions!"

We admire and reward…decisiveness, knowing the answer, and invention

+

Problems are uncomfortable and abstract

+

Our Brain is tricking us!

→ *Natural human tendency and bias towards jumping to solutions*

We subconsciously fill in the details we do not have and develop solutions.

Why is this a problem for AI implementations?

Presented with an AI solution end-users will reverse engineer what problem they *think* the solution was designed to solve and develop their own criteria for success and level of importance

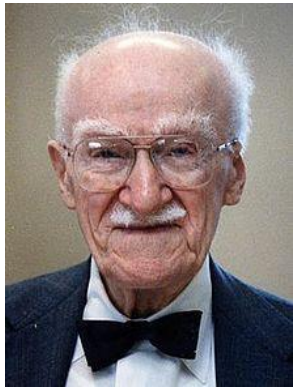
- Misinterpretation and misuse
- Disagreement on features and what is important
- Propagation of new skeptics
- Projects to fizzle ("the last mile problem")

"If I were given one hour to save the planet, I would spend 59 minutes defining the problem and one minute resolving it."

- Albert Einstein

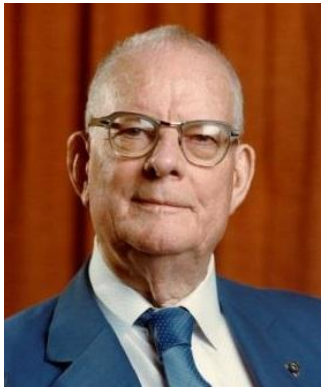
Process Improvement - Key Principles

Joseph M. Juran



Introduced the idea that **Diagnosis of the problem** must be done prior to Testing/Remediation and introduced the **Pareto Principle** to QI

W. Edwards Deming



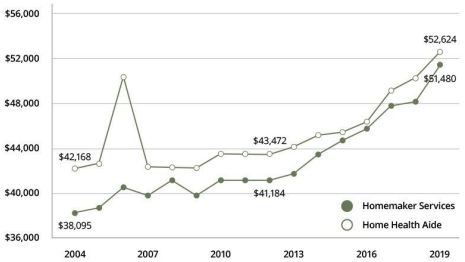
Introduced the System of Profound Knowledge facilitated by the **Plan, Do, Check, Act (PDCA)** Framework that promotes learning through testing.

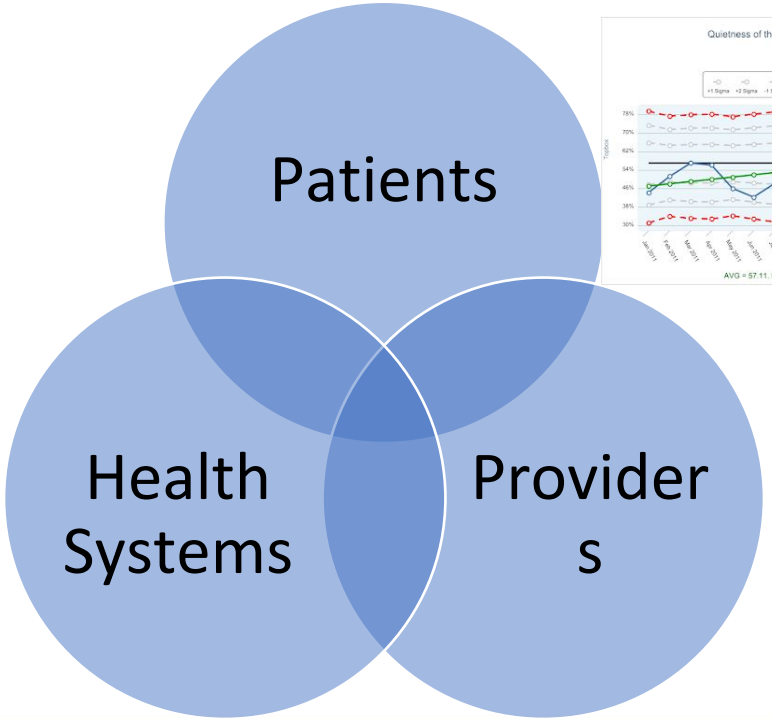
Problem Diagnosis Process

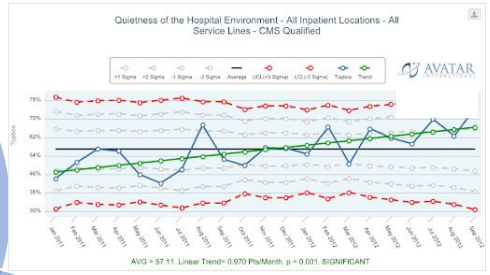
1. Identify a change in performance or sub-par performance
2. Map the critical processes yielding this performance
3. Obtain symptoms and discern the root causes
4. Prioritize and translate to key drivers for success
5. Assess utility of an AI-enabled solution
6. Design first pilot and conduct PDSA's

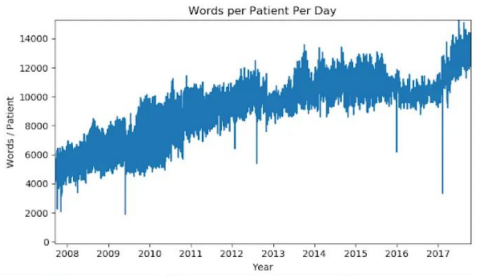
(1) Define the Problem

What is the measurable problem and who does it matter to?



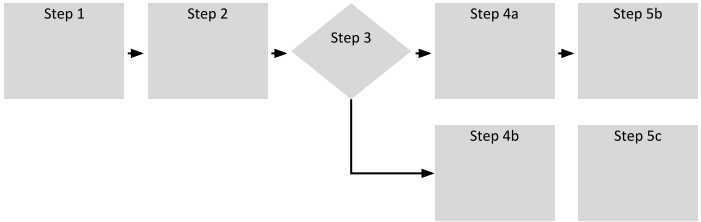






(2) Map the Critical Processes

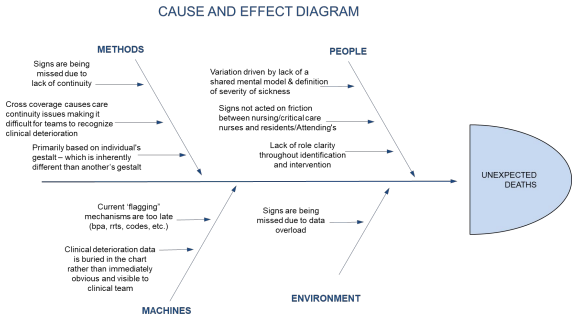
- Who are the actors?
- What tasks do they do?
- What steps do they take?
- What tools and technology do they use and when?
- What are they thinking? What decisions are made and why?



(3) Gathering Symptoms & Discerning Root Causes

Sociotechnical Systems Analysis

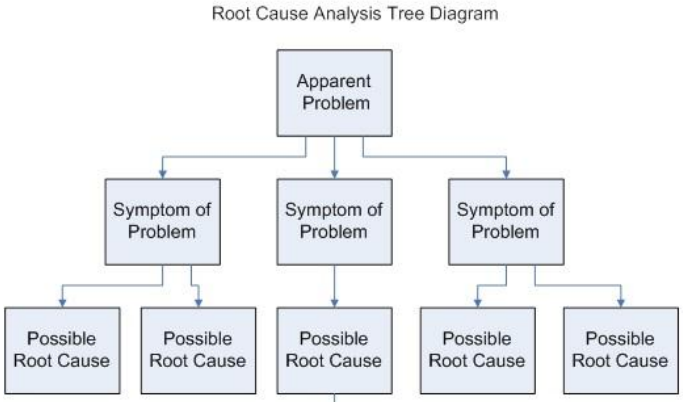
- People
- Process & Tasks
- Technology & Tools
- Internal Environment
- External Environment



Symptoms: observable evidence of the problem

- Errors
- Variation
- Friction between clinicians

Root Cause: Why?

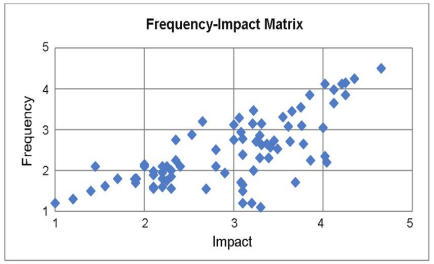


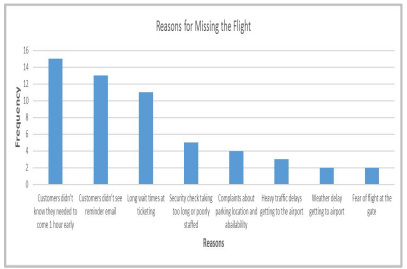
- Information loss between shift change

(4) Prioritize & Translate to Key Drivers

The **Pareto principle** states that for many outcomes roughly 80% of consequences come from 20% of the causes

Vilfredo Pareto





Key Drivers: describe the conditions that need to be true in order for performance to improve

- **Continuous** clinical status monitoring
- **Role clarity** throughout the process
- Clinical deterioration **detected early** with **clearly defined initial response and intervention**
- Agreed upon points when **appropriate core & extended team members will be involved**
- Agreed upon workflows **after initial response**
- **Objective** clinical assessment and shared mental model for risk of acute deterioration

Example:

Subjective clinical assessment with no share definition for risk of acute deterioration

(5) Assess the Utility of an AI-Enabled Solution

Classic Process Improvement Remediation Techniques:

- Eliminate wasteful activities
- Build consensus for appropriate standardization
- Minimize hand-offs
- Develop pull based processes reduce mental and physical inventory
- Design work cells

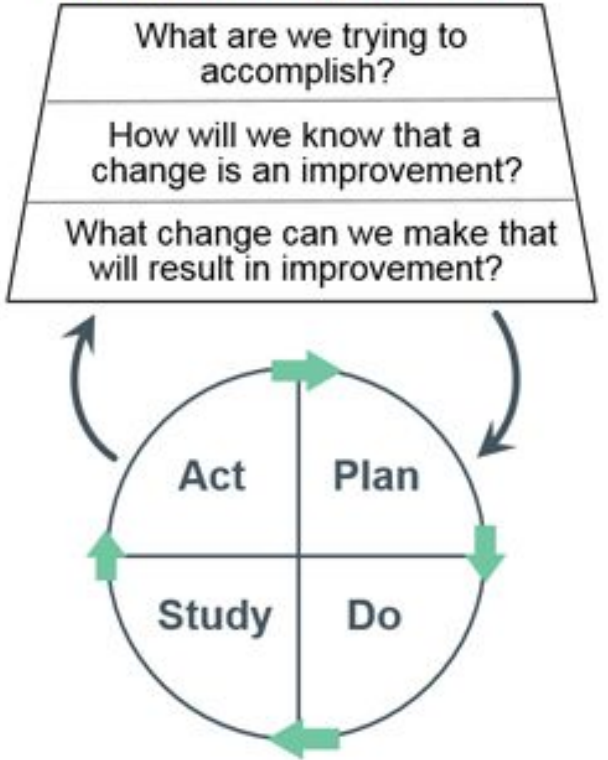
•Prediction and/or classification?

Continuous clinical status monitoring	Clinical deterioration **detected early** on wards	Consistent **identification and prioritization** of patients that need ACP

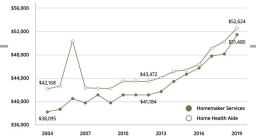
(6) Design and Pilot

- Conduct **future state process mapping sessions** leveraging design thinking methods and human factors engineering principles to stimulate creativity
  - *Anchor on your key drivers as 'guardrails' or 'design specifications'*
  - *Balance feasibility, acceptability, efficiency and effectiveness*

- Develop **prototypes**, and conduct body storming sessions using **real scenarios in a simulated environment**
  - *Allows for rapid testing of clinical integration workflows and AI model acceptability*



Recap



Identify sub-par performance

Map the critical processes

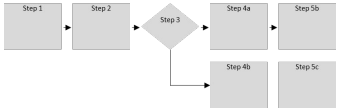
Discern the root causes

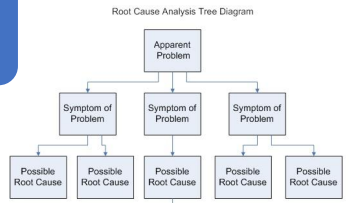
Prioritize and translate to Key Drivers

Assess utility of AI

Design and PDSA

Prevent jumping to AI-solutions





Continuous clinical status monitoring

Role clarity throughout the process

Clinical deterioration **detected early** with **clearly defined initial response and intervention**

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

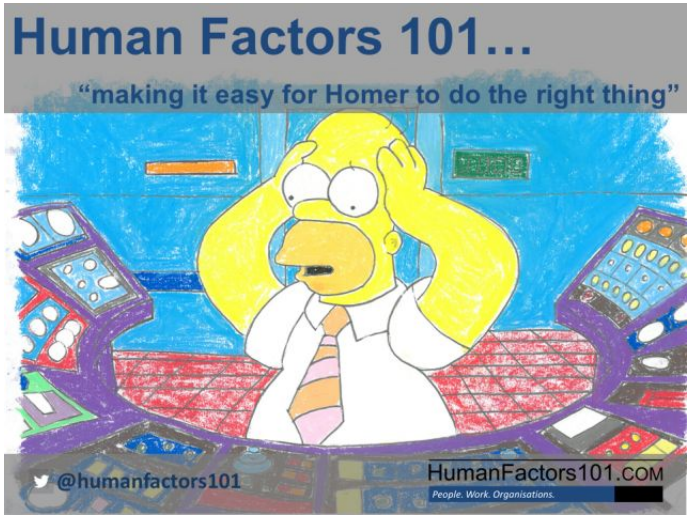
Human Factors Lens for AI Design, Implementation and Evaluation

Swaminathan Kandaswamy

Emory University

What is Human Factors?

- Multi Disciplinary Scientific Field
  - Including psychology, sociology, engineering, biomechanics, industrial design, physiology, anthropometry, interaction design, visual design, user experience, and user interface design
- studying human capabilities and understanding of interactions among humans and other elements of a system
- And applying theory, principles, data and methods to design of technology, systems, and processes in order to optimize human well-being and overall system performance for safety, efficiency, & quality.
- Used Extensively in other Industries Aviation, Automobile, Defense, Nuclear



Human Factors – Key Principle

- When introducing new design
  - We don’t redesign humans.
  - We redesign the system within which humans work.
- When automation is introduced
  - it transforms the tasks people do
  - It does not replace them

Humans and Automation

- How do we
  - increase appropriate use
  - decrease misuse and disuse
  - avoid abuse

Humans and Automation: Use, Misuse, Disuse, Abuse

RAJA PARASURAMAN,[1] *Catholic University of America, Washington, D.C., and* VICTOR RILEY, *Honeywell Technology Center, Minneapolis, Minnesota*

This paper addresses theoretical, empirical, and analytical studies pertaining to human use, misuse, disuse, and abuse of automation technology. *Use* refers to the voluntary activation or disengagement of automation by human operators. Trust, mental workload, and risk can influence automation use, but interactions between factors and large individual differences make prediction of automation use difficult. *Misuse* refers to overreliance on automation, which can result in failures of monitoring or decision biases. Factors affecting the monitoring of automation include workload, automation reliability and consistency, and the saliency of automation state indicators. *Disuse*, or the neglect or underutilization of automation, is commonly caused by alarms that activate falsely. This often occurs because the base rate of the condition to be detected is not considered in setting the trade-off between false alarms and omissions. Automation *abuse*, or the automation of functions by designers and implementation by managers without due regard for the consequences for human performance, tends to define the operator's roles as by-products of the automation. Automation abuse can also promote misuse and disuse of automation by human operators. Understanding the factors associated with each of these aspects of human use of automation can lead to improved system design, effective training methods, and judicious policies and procedures involving automation use.

How do we improve appropriate use, reduce misuse, disuse and abuse of AI?

- Define who does what- Define user, user roles and their tasks; Develop and use level of automation framework for Healthcare
- Define which aspect of HIP we are aiding – Understand AI capability and match requirements identified via observations, interviews and root cause analysis
- Consider context – Identify and understand contexts via observations, interviews and design interactions matching CCM
- Account for range of human behavior – Test choice for HAI by varying prominence, details of information, mode of recommendation(advice vs information) etc.
- Test design alternatives for performance tradeoffs- Performance, Trust (transparency and understandability), situational awareness, cognitive workload, motivation and skill degradation
- Long Term-Use – Monitor performance, Iterate and improve design based on feedback
- Avoid reinventing the wheel - Adopt learnings from literature and experience from other industries
- Crash test AI system before actual use!

Human Factors Lens for AI Design, Implementation and Evaluation

- Define “who does what?”
  - Identify User Roles and their Tasks – (AI Assistance is not necessarily always for a physician)
  - What is the role of clinician and what is the role of AI?
  - This essentially defines type and level of automation (LOA)

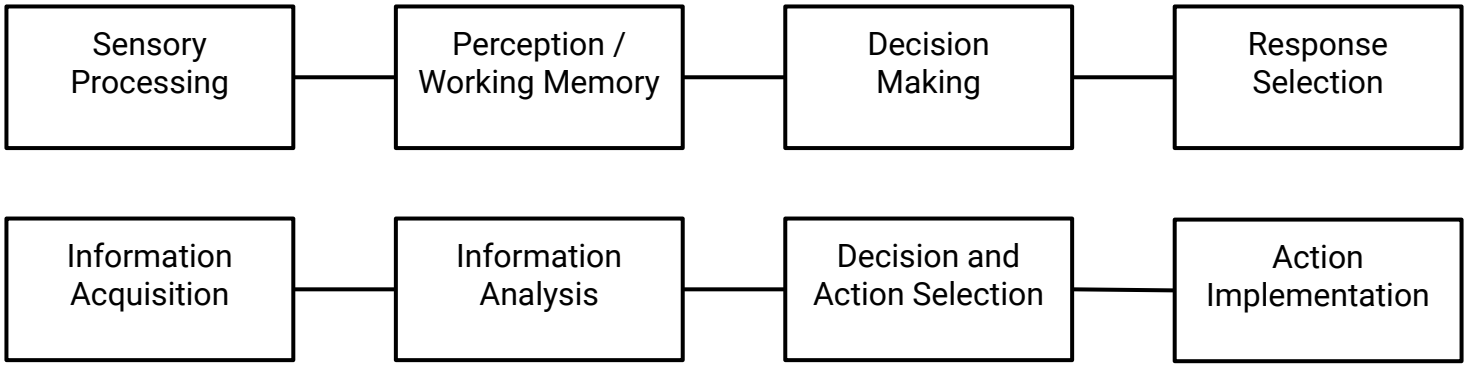
TABLE 1: Sheridan and Verplank (1978) Scale of Automation Levels

10. The computer decides everything and acts autonomously, ignoring the human.
9. ignores the human only if it, the computer, decides to
8. informs the human if asked, or
7. executes automatically, then necessarily informs the human, or
6. allows the human a restricted time to veto before automatic execution, or
5. executes the suggestion if the human approves, or
4. suggests one alternative, and
3. narrows the selection down to a few, or
2. The computer offers a complete set of decision/action alternatives, or
1. The computer offers no assistance: the human must take all decisions and actions.

Sheridan, T. B., Verplank, W. L. (1978). Human and computer control of undersea teleoperators.

Human Factors Lens for AI Design, Implementation and Evaluation

- Define which step(s) of Human Information Processing we are aiding automating with AI –
  Not always perception/ decision making
- Human Information Processing



Parasuraman, Raja & Sheridan, Thomas & Wickens, Christopher. (2000). A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 30(3), 286-297. IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society. 30. 286-97. 10.1109/3468.844354.

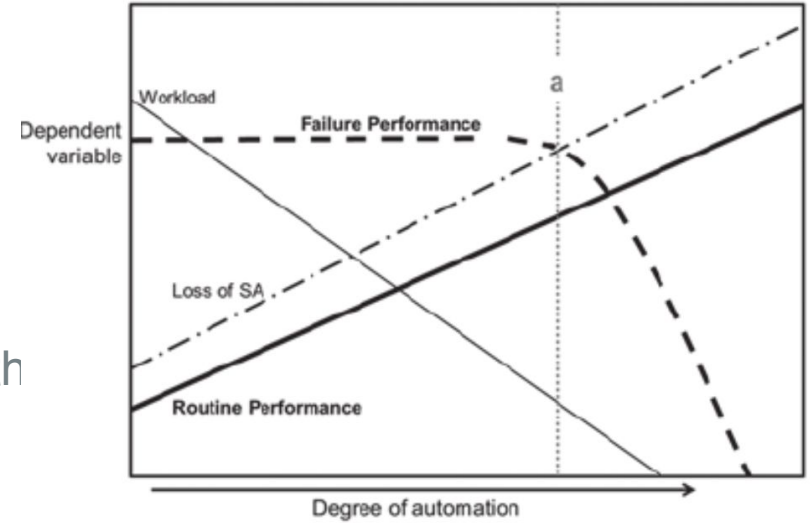
Human Factors Lens for AI Design, Implementation and Evaluation

Information Acquisition	filter (more, or less, “aggressively”) information from the EHR in support of clinician attention,
Information Analysis	integrate that information in a manner to form an assessment about patient state
Decision and Action Selection	recommend an action (labs, tests, monitoring, medications etc) to be taken based on the assessed in support of clinician's decision making, and
Action Implementation	carry out the physical action based on the recommended action in support of human muscular activity.

Wickens, Christopher. “Automation Stages & Levels, 20 Years After.” Journal of Cognitive Engineering and Decision Making, vol. 12, no. 1, Mar. 2018, pp. 35–41, doi:10.1177/1555343417727438.

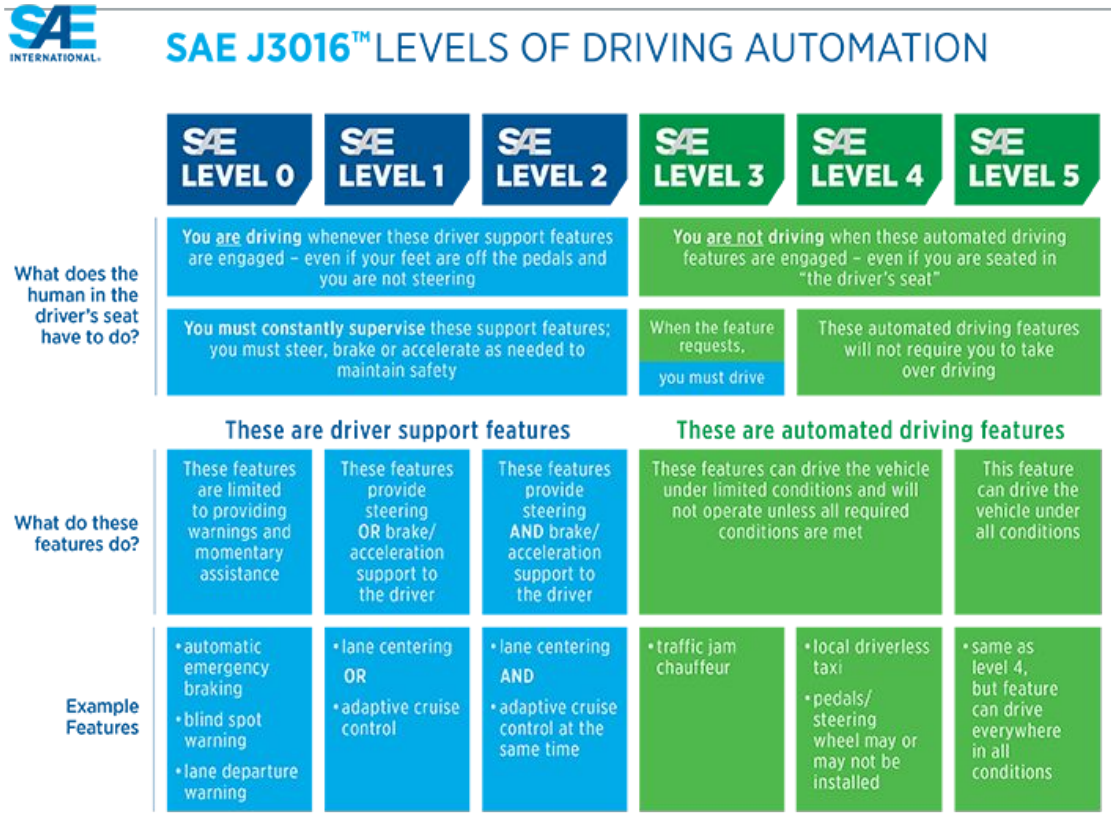
Human Factors Lens for AI Design, Implementation and Evaluation

- Why should we think about LOA?
  - For design specification - Functional allocation, roles and responsibility
  - Performance trade-offs
    - Increased performance and reduced workload with increasing automation
    - Degraded Situational Awareness with increasing automation
  - Prevent abuse of AI



Onnasch, Linda & Wickens, Christopher & Li, Huiyang & Manzey, Dietrich. (2013). Human Performance Consequences of Stages and Levels of Automation An Integrated Meta-Analysis. Human Factors The Journal of the Human Factors and Ergonomics Society. 10.1177/0018720813501549.

Human Factors Lens for AI Design, Implementation and Evaluation



- What does this look like for Healthcare?

Human Factors Lens for AI Design, Implementation and Evaluation

- Consider Context

- Strategic vs Tactical vs Opportunistic vs Scrambled
- With or Without Interruptions
- Individual vs Team
- Completing Alerts

Human Factors Lens for AI Design, Implementation and Evaluation

- Define and evaluate design choices for trade off between important human performance outcomes
  - Cognitive workload
  - Trust
  - Situational Awareness
  - Motivation
  - Skill degradation

Human Factors Lens for AI Design, Implementation and Evaluation

- Users will not always do what designers expect them to do
  - Automation Bias - propensity for humans to favor suggestions
  - Complacency – satisfaction with current solution but may lack awareness of other safer or more efficient options
  - Satisficing- behavior to accept most accessible solution that meets minimal level of performance, often relying on use of the heuristic to reduce cognitive load and speed-up performance
  - Vigilance decrements
- These result in misuse and disuse of AI

- How do we evaluate human performance?
- How do we account for a range of human behavior?

Kate Goddard, Abdul Roudsari, Jeremy C Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, *Journal of the American Medical Informatics Association*, Volume 19, Issue 1, January 2012, Pages 121–127, https://doi.org/10.1136/amiajnl-2011-000089

Human Factors Lens for AI Design, Implementation and Evaluation

• Expanding design questions

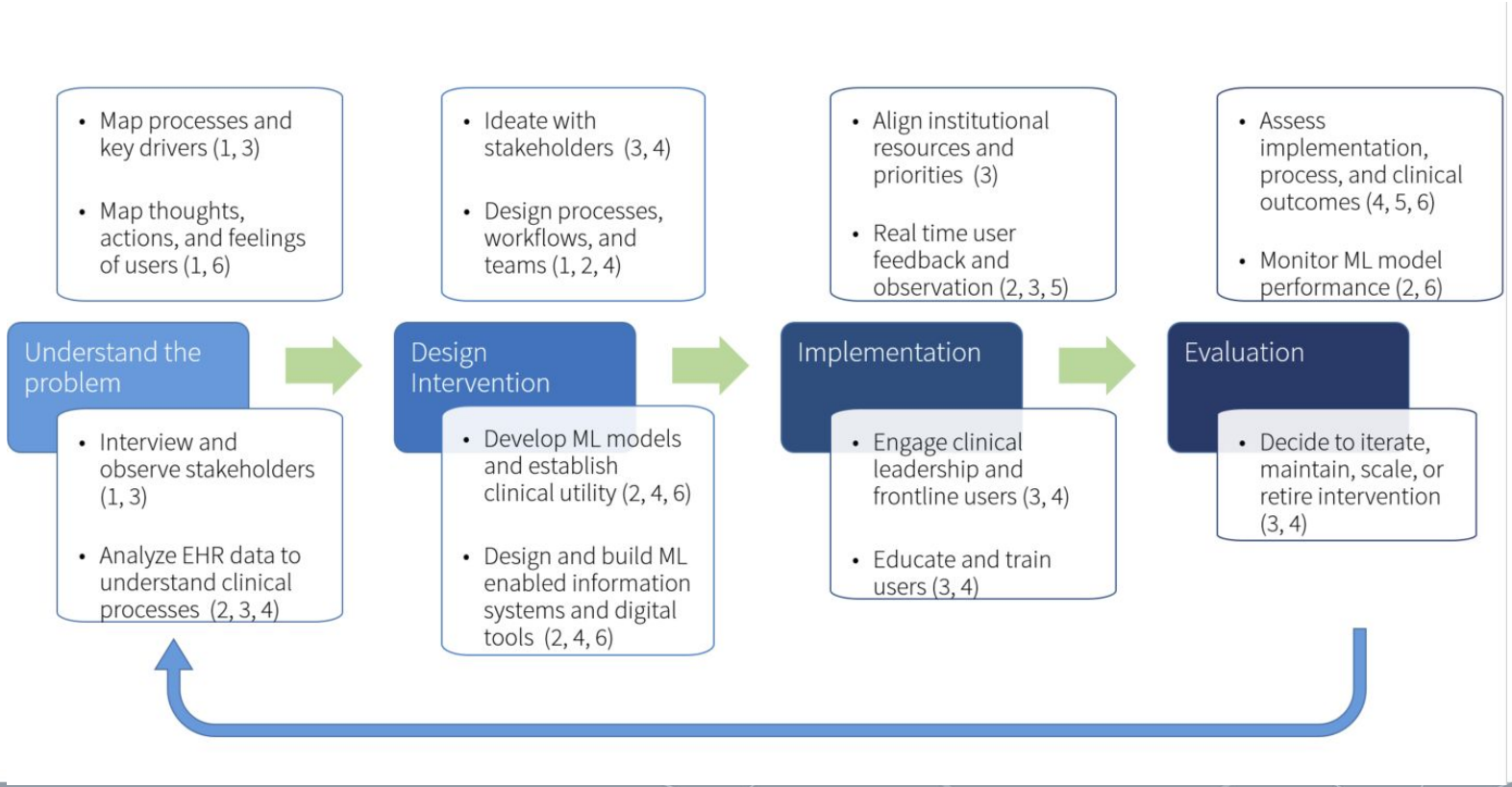
- How implementing AI models is going to change social systems?
- How social systems are going to change use of AI?
- What is the range of scenarios that need to be considered?
- What interactions could / should be supported?
- How will clinician perform in scenarios that exceed the competence limits of AI?

Smith, Philip J. “Conceptual Frameworks to Guide Design.” Journal of Cognitive Engineering and Decision Making, vol. 12, no. 1, Mar. 2018, pp. 50–52, doi:10.1177/1555343417732239.

Recap - How do we improve appropriate use, reduce misuse, disuse & abuse of AI?

- Define who does what- Define user, user roles and their tasks; Develop and use level of automation framework for Healthcare
- Define which aspect of HIP we are aiding – Understand AI capability and match requirements identified via observations, interviews and root cause analysis
- Account for context – Identify and understand contexts via observations, interviews and design interactions matching CCM
- Account for range of human behavior – Test choice for HAI by varying prominence, details of information, mode of recommendation(advice vs information) etc.
- Test design alternatives for performance tradeoffs- Performance, Trust (transparency and understandability), situational awareness, cognitive workload, motivation and skill degradation
- Long Term-Use – Monitor performance, Iterate and improve design based on feedback
- Avoid reinventing the wheel - Adopt learnings from literature and experience from other industries
- Crash test AI system before actual use!

A multidisciplinary process for developing and implementing AI



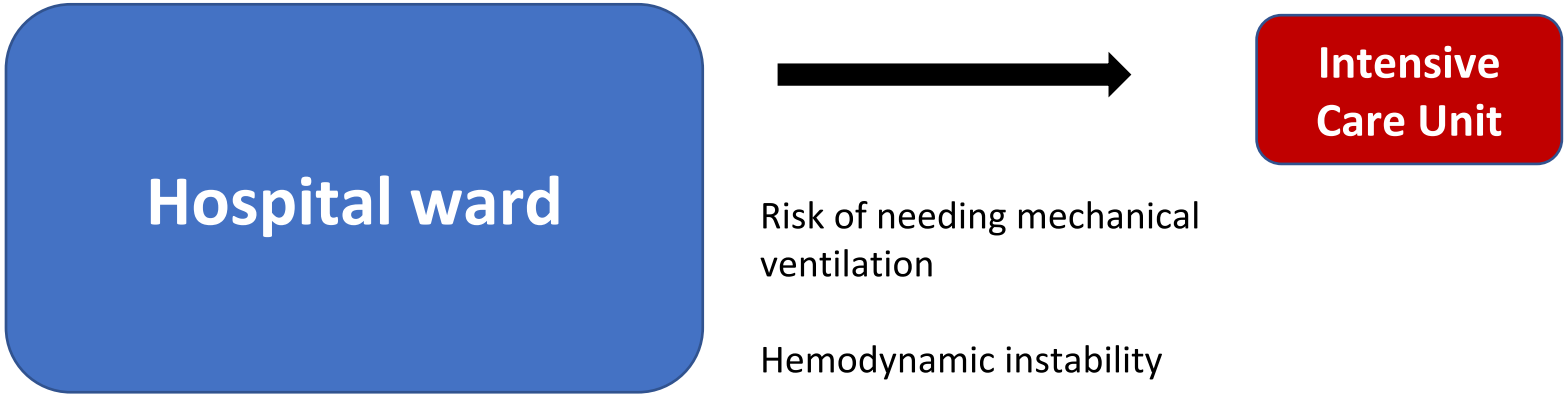
Li, R.C., Asch, S.M. & Shah, N.H. Developing a delivery science for artificial intelligence in healthcare. *npj Digit. Med.* **3,** 107 (2020).

Real world examples at two academic medical institutions

Stanford Hospital is a 605 bed adult tertiary care center in Palo Alto, California.

Children's Hospital of Philadelphia (CHOP) is 559 bed pediatric tertiary care hospital and also the community hospital for West Philadelphia.

How can we identify sick patients early before they emergently need critical care?



1. Understanding the problem

The director hospital quality tells you that the hospital would like to lower inpatient morbidity and mortality, and they want to leverage informatics and AI to solve this problem.

How can we better define and understand this problem?

How do we know if/how machine learning should be part of the solution?

Early recognition of pediatric deterioration

- 2017 study: 215 children from 23 hospitals who required CPR on the wards or within 48 hours of transfer to the ICU, only 22% had a preceding rapid response team (RRT) evaluation. RRTs known to be effective as well.
- Every critical deterioration event (ICU transfer requiring pressors or positive pressure within 12 hours) estimated to add nearly $100,000 to the cost of hospitalization at CHOP and review at other institutions suggest that more than 40% of events may be preventable

Defining the Problem

Initially formed a workgroup for FY20 to assess decision to renew / “replace” vendor EWS

Preventing Codes Outside the ICU (PCOTI) Harm Prevention Program, rolled out new Critical Care Outreach Team and tiered review process for ICU transfers

Redefining the systems problem

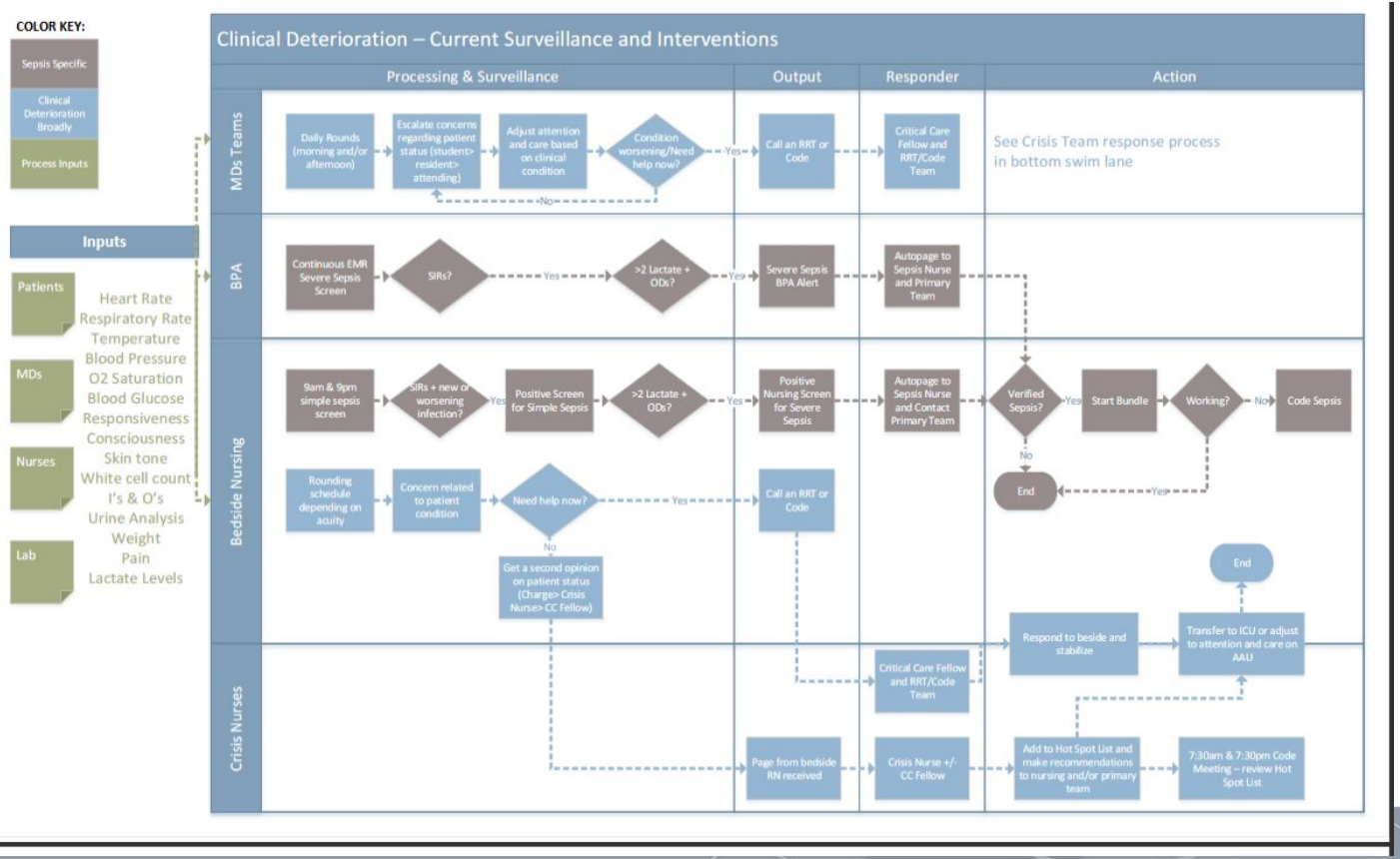
Improve the rate of the CHOP early recognition system (i.e. people, process, technology) identifying children in the 2-24 hours prior to deterioration and engage critical care resources to mitigate risk

Ultimately, reduce the rate of emergency transfers and codes outside the ICUs

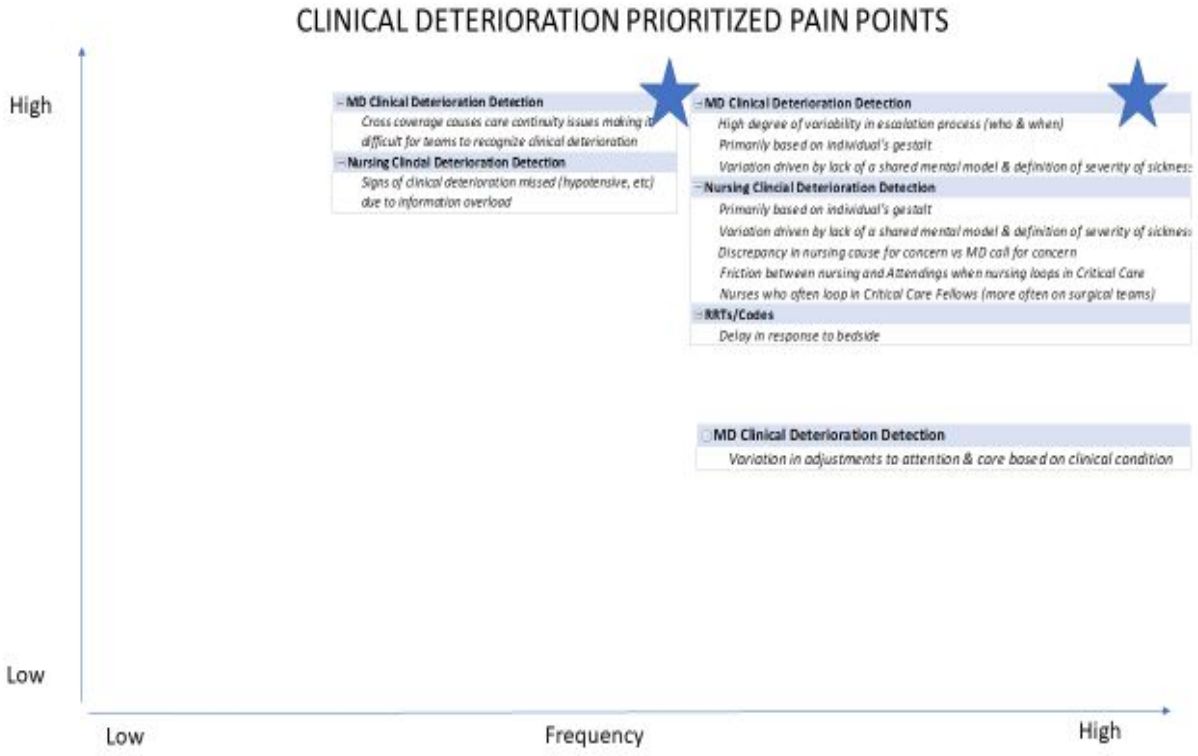
Understanding the current state

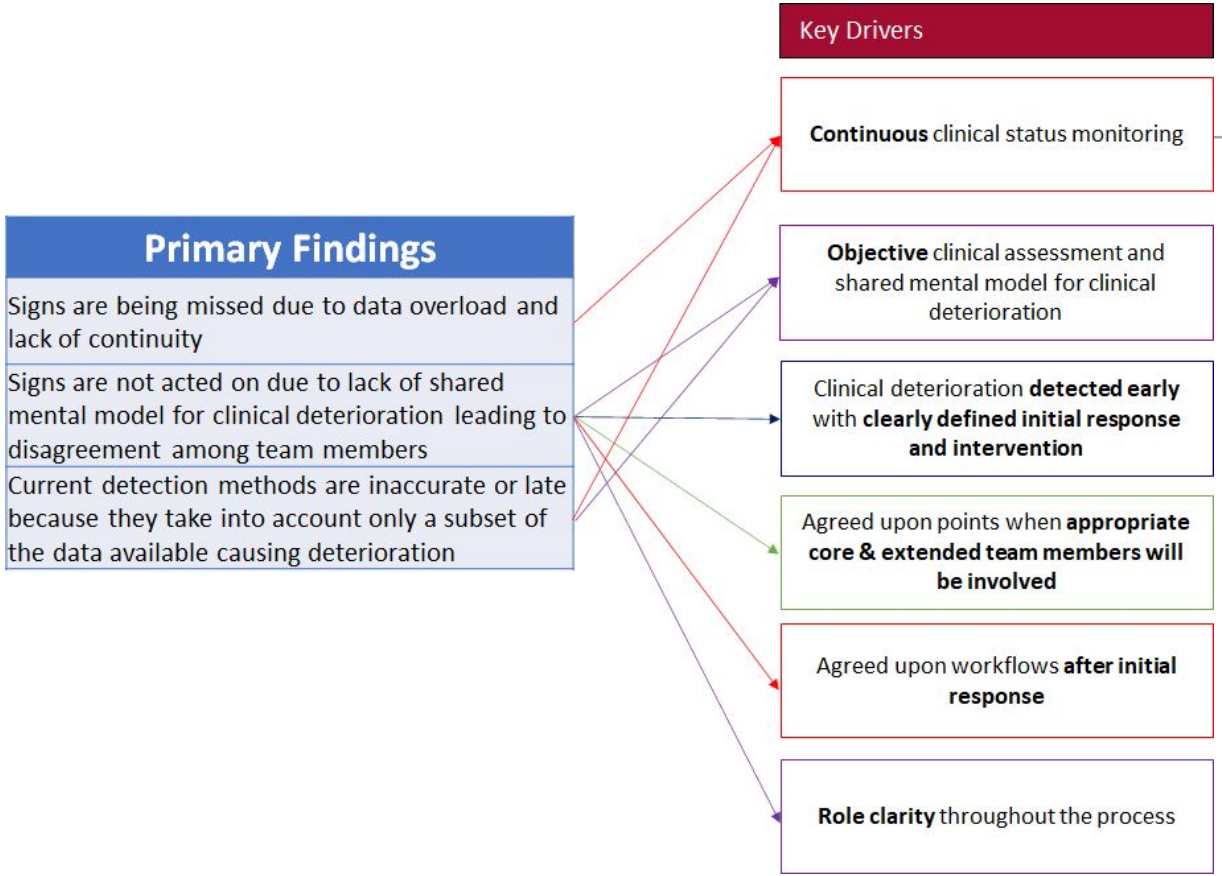
Children's Hospital of Philadelphia

Process mapping of current state workflows



Focus on highest yield pain points





Derive key drivers for an improved work system

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

Stanford
HEALTH CARE
STANFORD MEDICINE

Designing the solution

Based on the pain points and key drivers identified by the team …

How can these pain points be translated into features of a solution? What are the key components of such a solution?

How does machine learning fit into such a solution? How should the machine learning task be identified? How to select the best model?

What human factors concepts do we need to consider when designing the solution?

Creating an AI-enabled work system

Stanford HEALTH CARE
STANFORD MEDICINE

Within 60 mins

Key Drivers

- **Continuous** clinical status monitoring
- **Objective** clinical assessment and shared mental model for clinical deterioration
- Clinical deterioration **detected early** with **clearly defined initial response and intervention**
- Agreed upon points when **appropriate core & extended team members will be involved**
- Agreed upon workflows **after initial response**
- **Role clarity** throughout the process

Patient Status

CDI Model running on **C3 patients**; recalculating score every 15 minutes

Patient reach >= 65; **1/5 chance of RRT or ICU transfer in 6-18 hours**

Shared Mental Model/Awareness

Alert **Primary Team MD** via Spok

Automatically added to shared patient list **visible to CC Triage & CCRNs**

Non-interruptive BPA **visible to bedside nursing and charge nurse** and remain on landing page

Action and Follow-Thru

Voalte Bedside RN Do we need to evaluate at the beside?

What are you **worried** about? CC fellow **consult?** Scenario **planning**? **Care plan** adjustments?

Bedside RN page CCRN with concerns about deterioration (SBAR) & contact resident

Bedside RN connect with Charge Nurse

Documentation (details TBD)

Patient Status

If a patient is >= 65

Shared Mental Model/Awareness

Non-interruptive BPA **remains visible to all** on landing page & wrenched in column

Action and Follow-Thru

MDs & nursing **huddle at bedside rounds** > scenario planning and care plan adjustments

Document (details TBD)

Identifying the prediction task and model

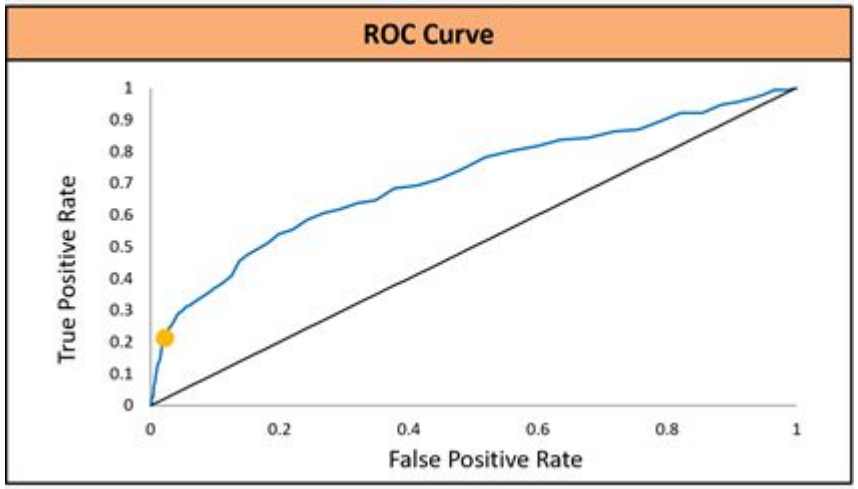
Prediction task requirements

- Outcome should reflect an overall state of “being critically ill”
- Time of prediction needs to be early enough prior to the time of outcome

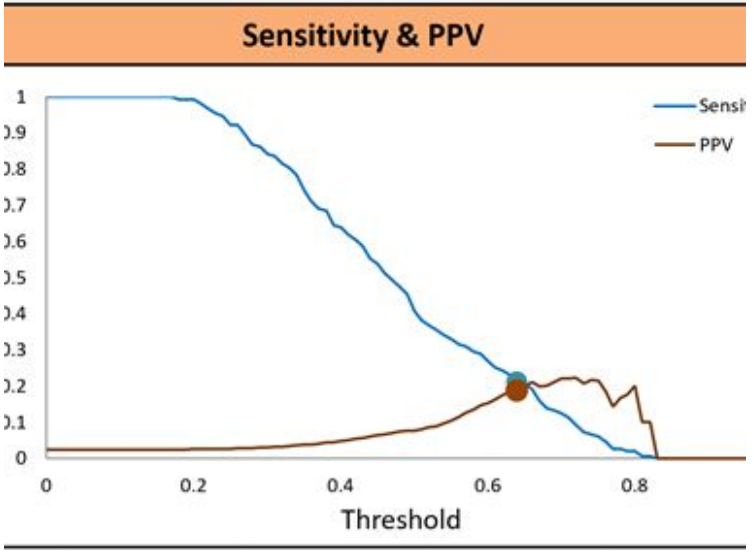
Model selection

- Model developed by EHR vendor (Epic Systems)
- Trained across three hospitals with ~327k patient encounters using ordinal labels (rapid response/code event, ICU transfer, inpatient death)
- At runtime, outputs a “deterioration index” from 0-100

Optimizing the model threshold for the clinical use case

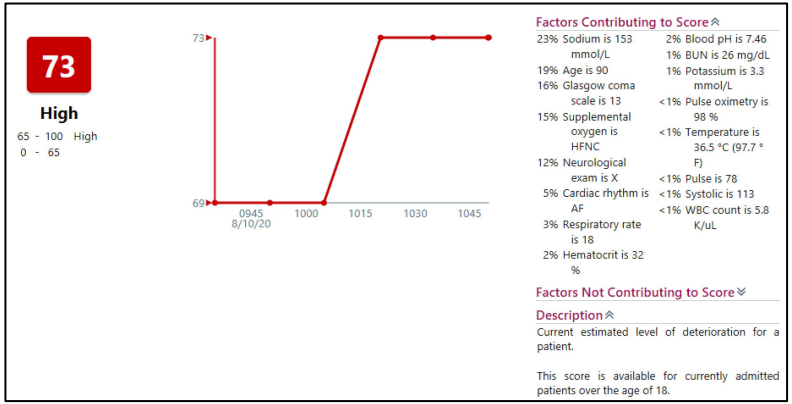


Model validated over 6k inpatient encounters to assess performance for predicting events in **6-18 hours**



Threshold of 65 chosen to achieve 20% PPV and recall (estimated ~2 new alerts per day on medicine service)

Admission Date	Diagnosis	Risk of Clinical Deterioration
8/5/20	Vomiting and diarrhea	.....
8/8/20	Alcohol-indu… acute pancreatitis,…	.....
8/4/20	Hypervolemia, unspecified hypervolemi…	.....
8/8/20	Bacterial urinary infection	.....
8/6/20	Acute respiratory failure with…	!!
7/17/20	CKD (chronic kidney disease) stage 4, GF…	!!
8/7/20	Acute on chronic heart failure,…	.....
8/10/20	Cellulitis, unspecified	.....



Designing an interface that integrates ML model predictions with clinical decision making and workflow

Stanford HEALTH CARE
STANFORD MEDICINE

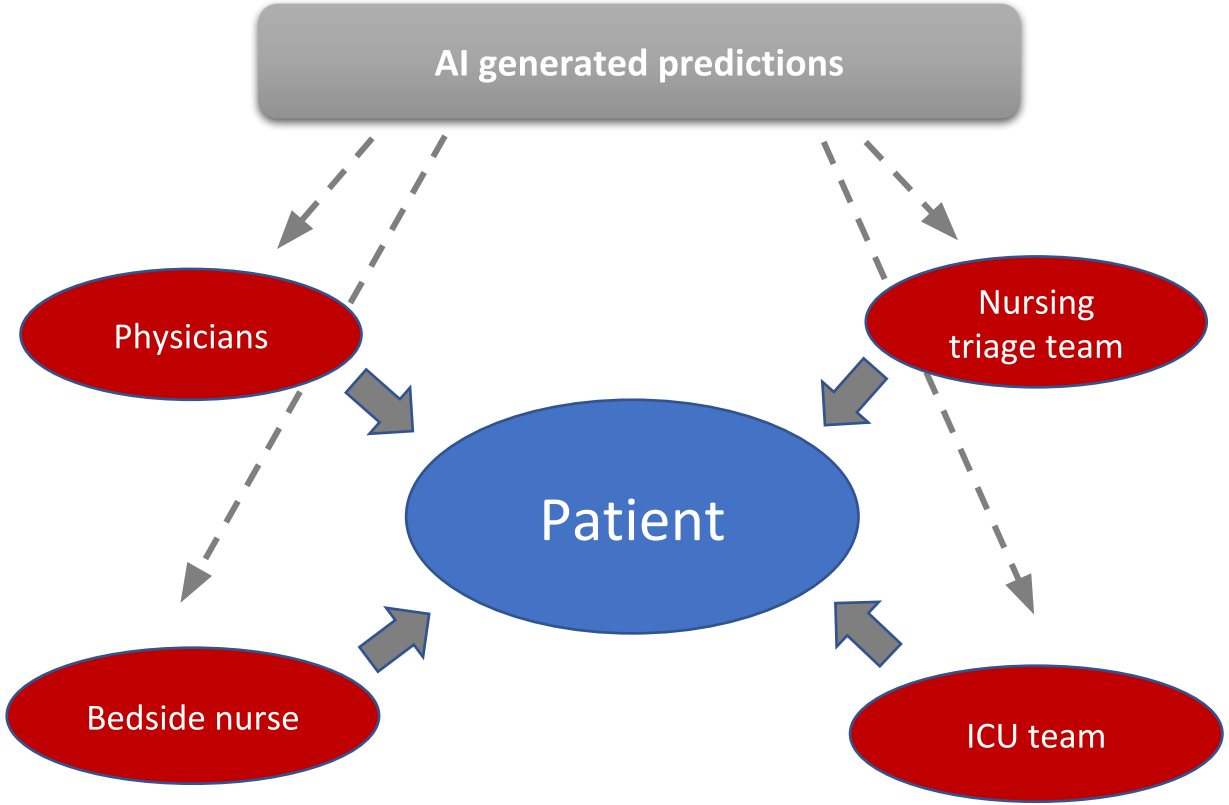
Team based checklist for patients flagged by the model

Situation
Patient flagged by ML model
Background
Airway/Oxygenation needs: [free text]
IV access: [free text]
Likely reason for deterioration: [free text]
Assessment
Vital signs [auto populate]
Free text any other pertinent information (mental status, med admin info ie last time pain med given)
Response
Interventions:
- Changes to care management (e.g. Diagnostics, antibiotics, etc) [free text]
- Aspiration precautions [discrete bundle]
- Critical Care Consult [yes/no?]

Scenario Planning
- What steps do we need to take if the patient continues to deteriorate? [free text]
- Code status/Goals of Care (e.g. ok to escalate to ICU?) […]

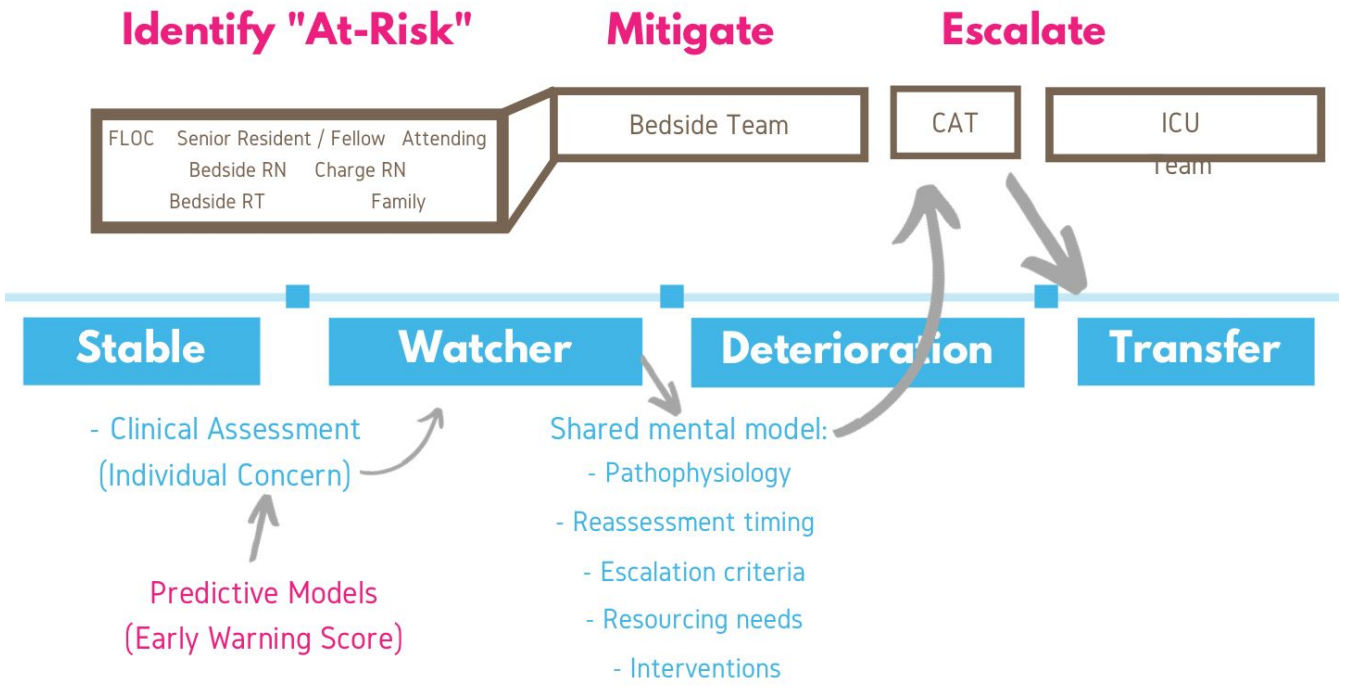
Stanford HEALTH CARE
STANFORD MEDICINE

An intelligent system that enables new team structures and roles



Stanford HEALTH CARE
STANFORD MEDICINE

Defining the process and prediction task



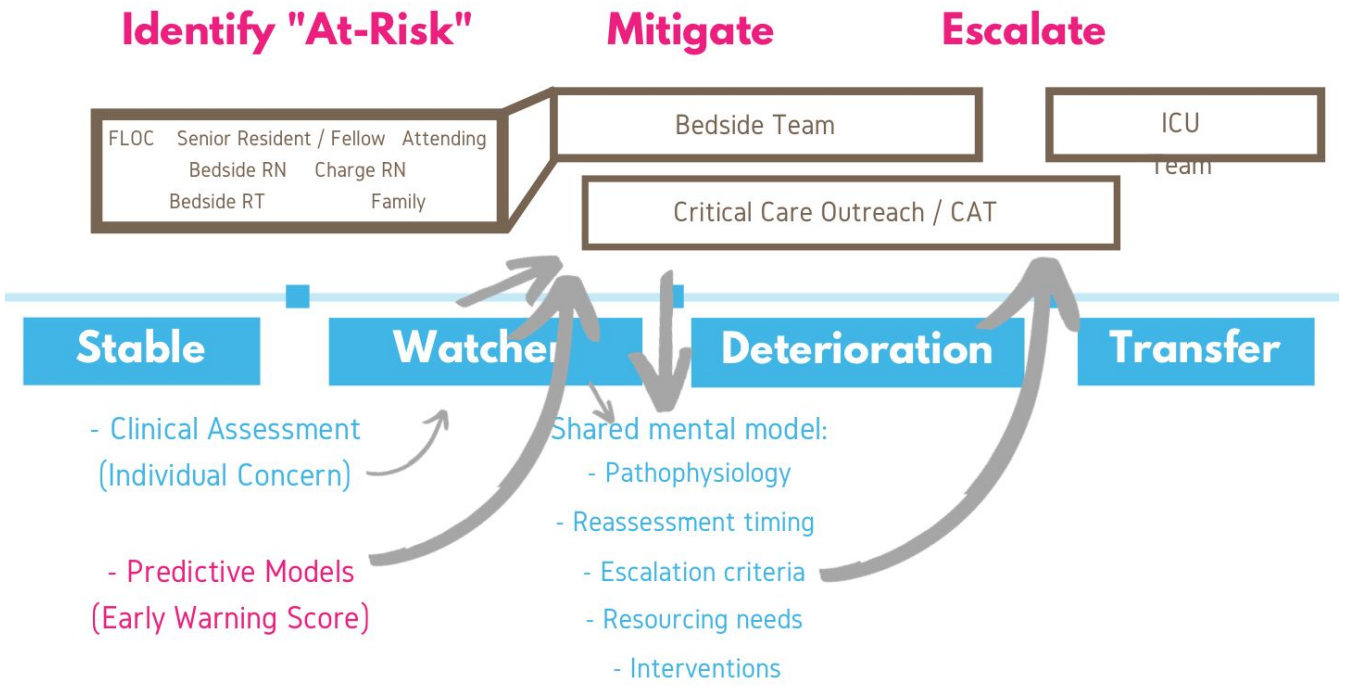
Beyond Reporting Early Warning Score Sensitivity: The Temporal Relationship and Clinical Relevance of "True Positive" Alerts that Precede Critical Deterioration

Meredith C Winter, MD[1]*, Sherri Kubis, RN, BSN, CCRN[2], Christopher P Bonafide, MD, MSCE[3,4,5]

The main finding of this study is that 90% of CDE events that generated "true positive" alerts had evidence suggesting that clinicians had already recognized deterioration and/ or were already escalating care before most alerts would have been triggered.

Other aspect is what does the model add in addition to the human Watcher process – surprisingly little overlap between EWS alerts and team designation of high risk patients of high risk patients.

Defining the process and prediction task



Comparing prediction models / design

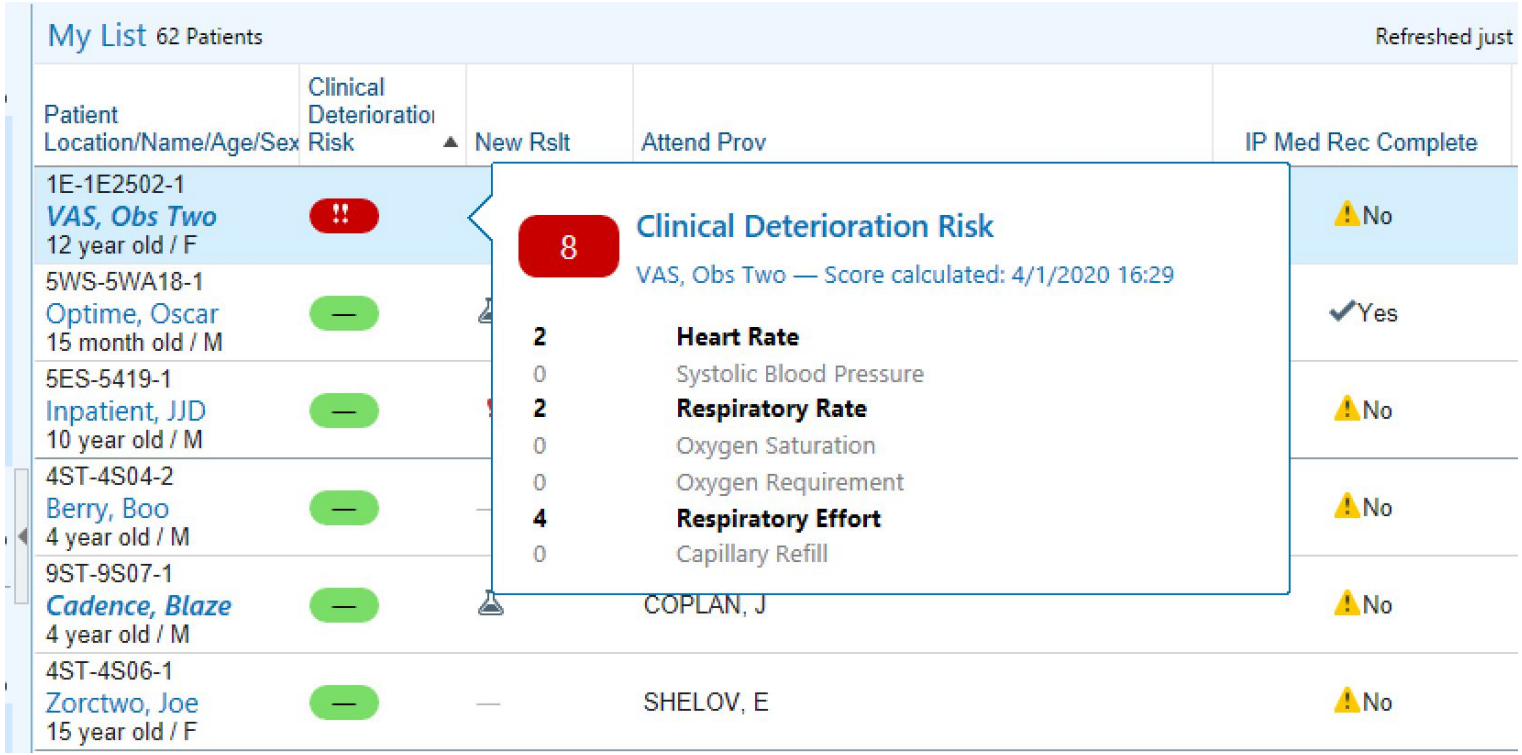
Traditionally, model comparison based on AUC curve, but is an improvement in AUC clinically significant?

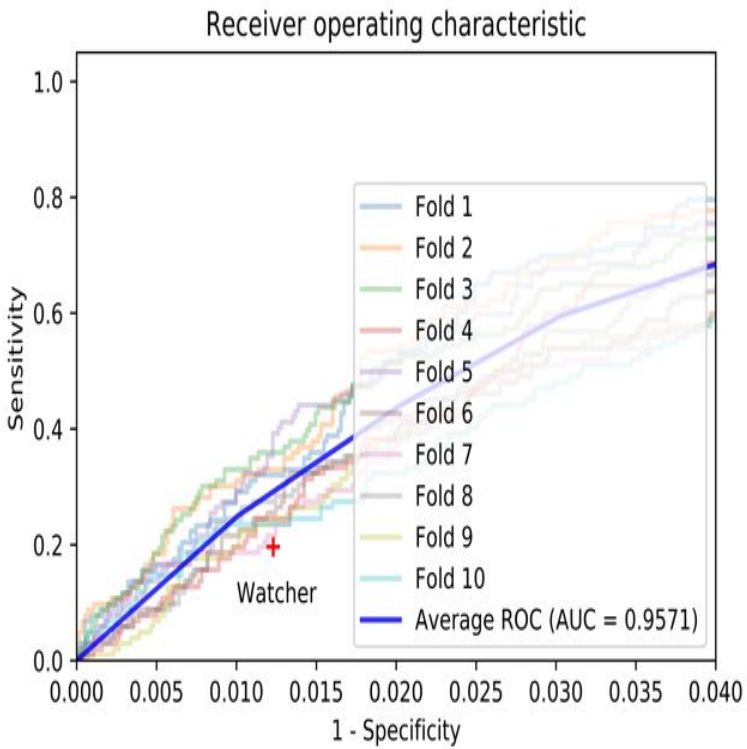
- Vendor high-risk alerts: ~25% PPV, ~30% sensitivity
- BedsidePEWS: ~15% PPV, ~55% sensitivity
- BedsidePEWS preferred – free, CCOT ok with >10% PPV

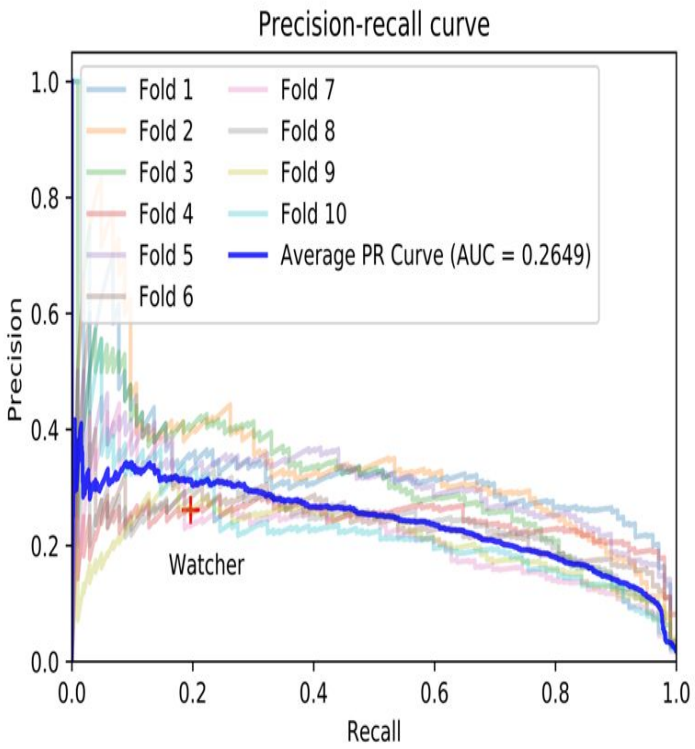
One solution: could have adjusted threshold for vendor model.

We ultimately stuck with BedsidePEWS because of cost.

Workflow integration







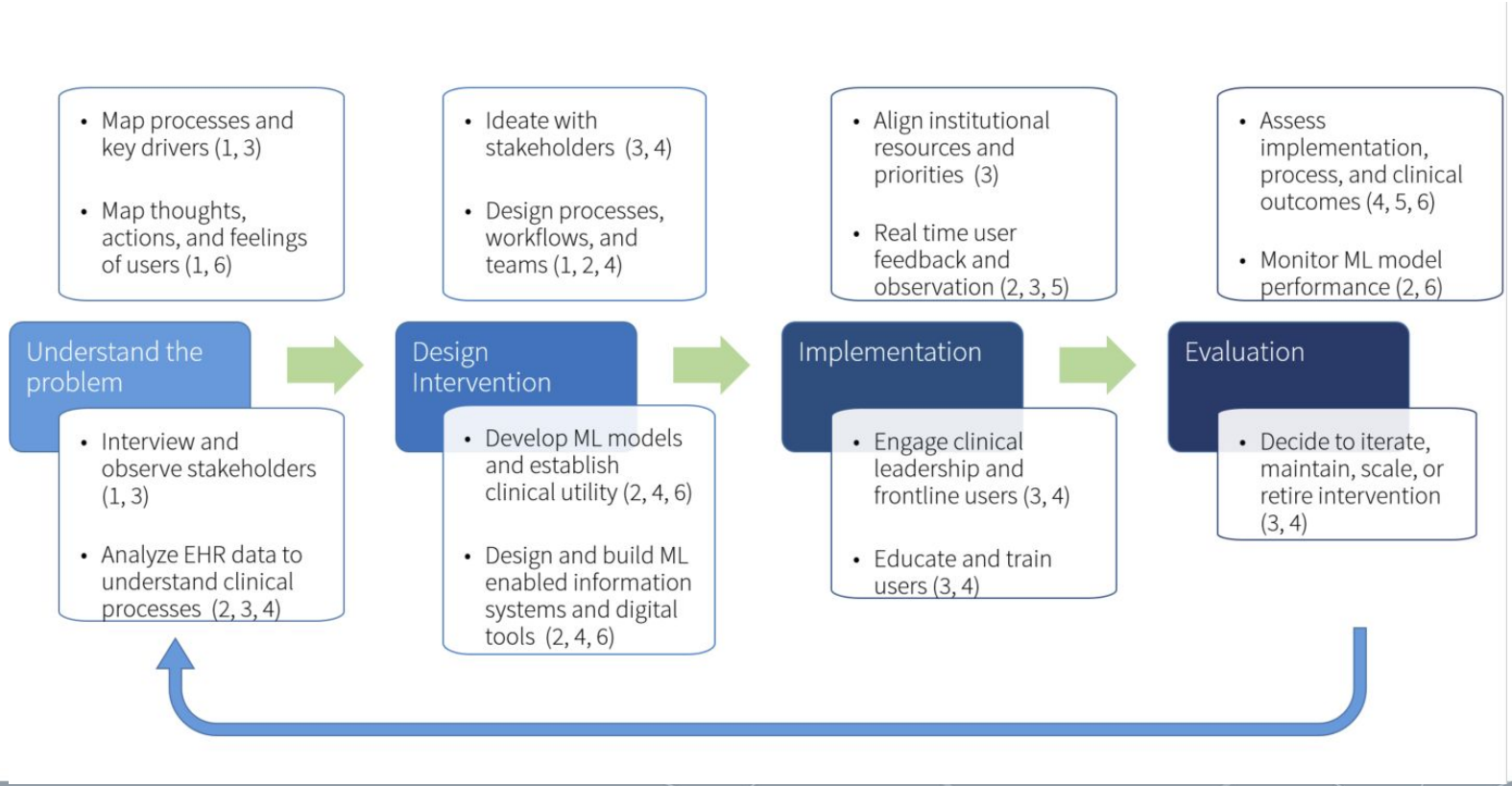
Rich Tsui, Helen Shi, Jerry Shaeffer

Evaluation of models

- With our advanced model, may not matter – may be good enough on its own (as opposed to prediction + Watcher) though still may not be helpful for bedside based on the PPV of ~30% for “highest risk.”

- Now initiated study on the effectiveness of our models on our aim of early targeting of Critical Care Outreach Team.

A multidisciplinary approach towards using AI to solve real world problems in healthcare



Li, R.C., Asch, S.M. & Shah, N.H. Developing a delivery science for artificial intelligence in healthcare. *npj Digit. Med.* **3,** 107 (2020).

Thank You!

Ron Li, MD

ronl@stanford.edu

Twitter: @ronlivs

Naveen Muthu, MD

muthun@chop.edu

Twitter: @naveenmuthu

Swaminathan Kandaswamy, PhD

swaminathan.kandaswamy@emory.edu
Twitter: @swamikandaswamy

Jonathan H. Chen, MD PhD

jonc101@stanford.edu

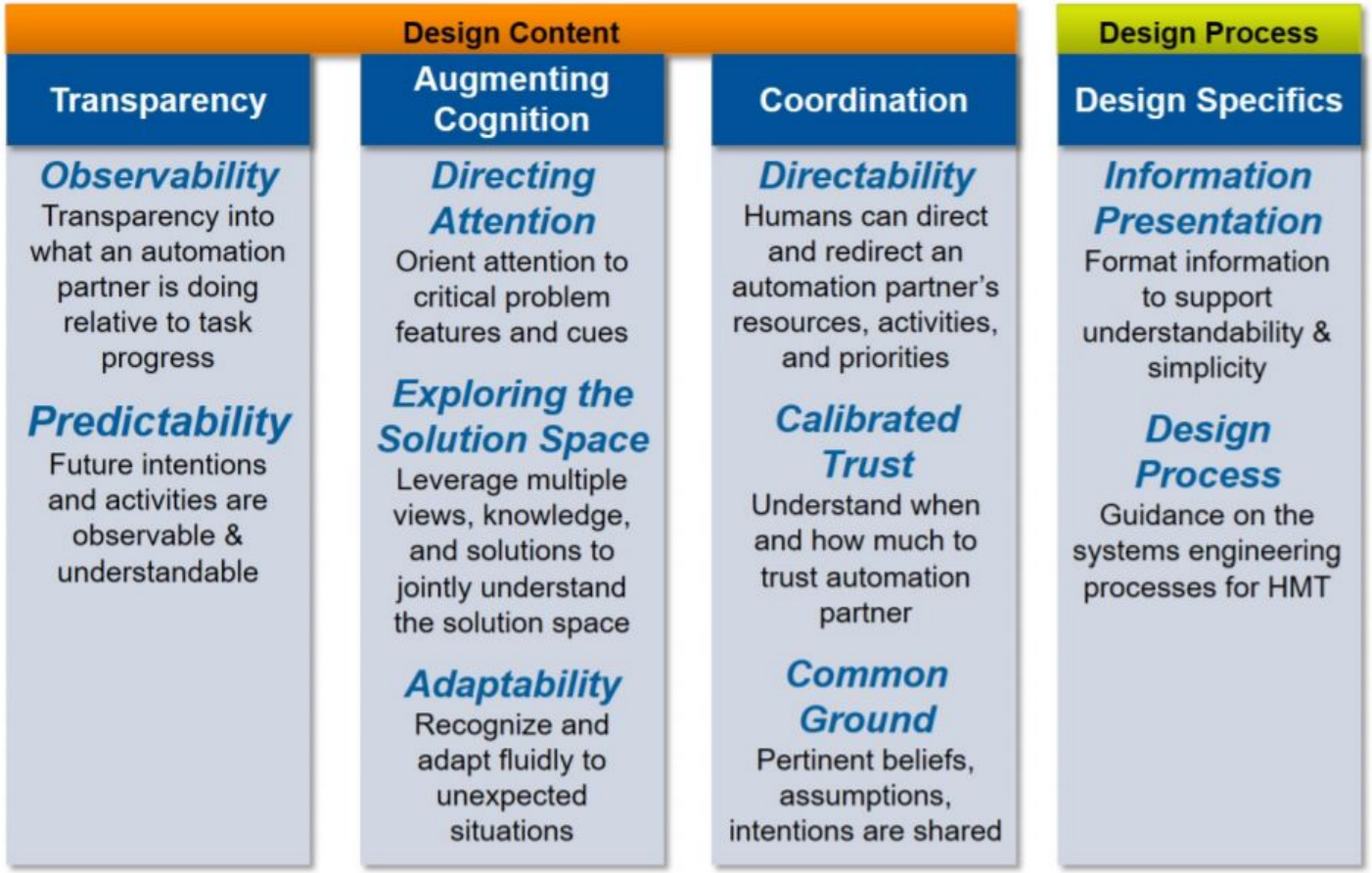
Twitter: @jonc101x

Margaret Smith, MBA

Marsmith@stanford.edu
med.stanford.edu/healthcare-ai

Appendix

Some Guidelines

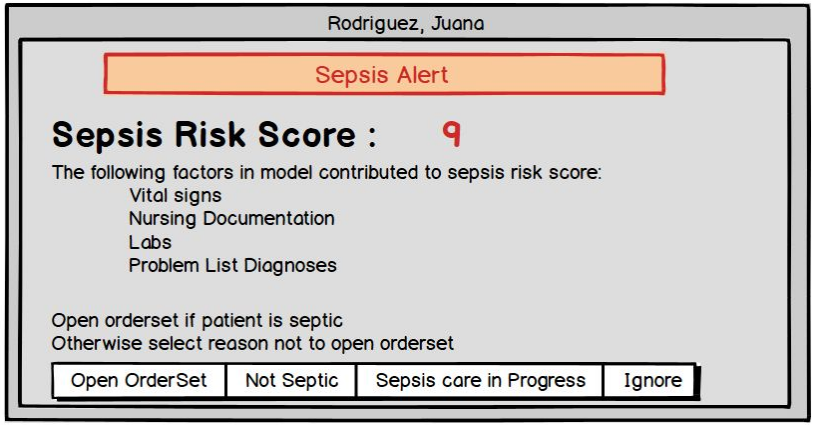


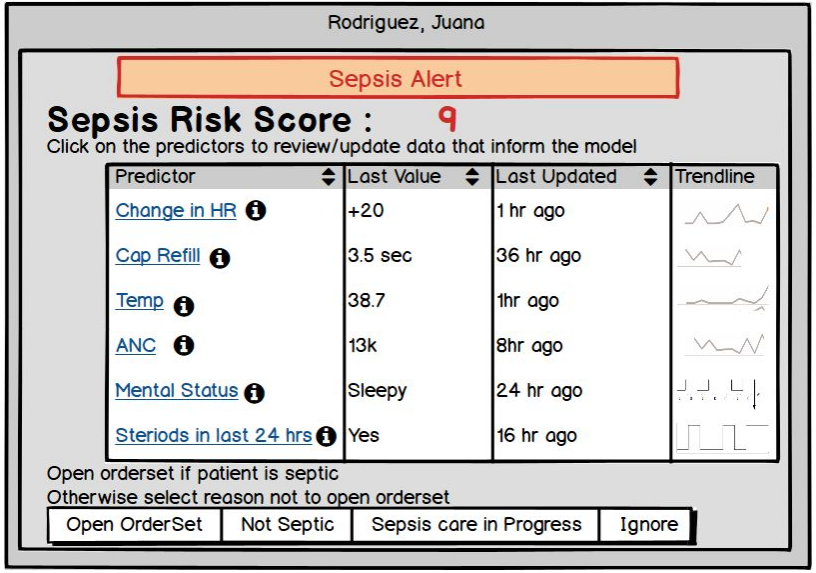
P. Mcdermott, C. Dominguez, N. Kasdaglis, M. Ryan, I. T. Mitre, and A. Nelson, "Human-Machine Teaming Systems Engineering Guide," 2018.

Some Guidelines

- Observability - Why did the model make this decision?
- Observability , Predictability - How confident is the model? And why is it confident?
  - Model should show low confidence when its prediction can be wrong
- Calibrate Trust- Help clinicians provide feedback to increase confidence?
- Calibrate Trust- ,Common Ground, Information Presentation - Help clinician know when model can fail, convey model assumptions in a way that is
  - noticeable to clinicians
  - understandable i.e words, terms that are familiar to them not just model developers.
- Direct Attention- Tell clinician what factors to consider that they are not viewing
- Explore solution space- Direct clinicians to what they can do/ diagnosis alternatives that they are not already doing/considering?

Some Guidelines





Some Guidelines

- Complacency error - can be reduced by training
- Automation Bias (AB)
  - Increasing accountability for decisions may reduce AB
  - Display prominence increases AB - prominent incorrect advice is more likely to be followed
  - Too much on-screen detail can increase biases as it makes people less conservative
  - Mode of advice (information vs recommendation)
  - Presentation of additional information such as when last change was made model sensitivity to change in input etc. can improve appropriate reliance
- Vigilance
  - Varying reliability can increase vigilant behavior

Kate Goddard, Abdul Roudsari, Jeremy C Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, *Journal of the American Medical Informatics Association*, Volume 19, Issue 1, January 2012, Pages 121–127, https://doi.org/10.1136/amiajnl-2011-000089

Kaber, David B. "Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation." Journal of Cognitive Engineering and Decision Making, vol. 12, no. 1, Mar. 2018, pp. 7–24, doi:10.1177/1555343417737203.

Some Guidelines

- Accounting for nonstandard behaviors – design to reduce user effort
  - Identify and evaluate system outcomes for different Cognitive Control Modes
    - Strategic mode – user looks at looks ahead at higher level goals, will have evaluated the outcome more precisely, and considered the relationship between action and its pre-conditions; o time constraint
    - Tactical mode - pre-planned action, where the user will use known rules and procedures to plan and carry out short term actions; but still is under time constraint
    - Opportunistic mode - chance action taken due to time, constraints and again lack of knowledge or expertise and an abnormal environmental state; users revert to using heuristics
    - Scrambled mode - completely unpredictable situation where the user has no control and must act in an unplanned manner
  - Design interfaces matching different CCMs to effectively support assessments
- Example form Aviation
  - Design of decision support for airline operational managers
  - Design enabled decision maker to select different interface modes during system operation that could support their CCM.
  - Study found superior system performance when the interface mode mirrored the CCM.
  - Design disrupted cognitive activity in contextual control modes which they were not intended to support

Feigh, K. M. (2011). Incorporating multiple patterns of activity into the design of cognitive work support systems. Cognition, Technology & Work, 13(4), 259–279
Hollnagel, Erik. (1998). Context, cognition, and control. Co-Operation in Process Management-Cognition and Information Technology..

Some Guidelines

Four Life Cycle Phases of Artificial Intelligence Model Deployment Incorporating Human Factors Elements and User-centered Design

Life Cycle Phase	Description	Example User-Centered Design Methods and Techniques
Design	The intended user of the AI should be involved early and continuously during this stage to ensure their needs are considered.	Observe the clinical environment Identify needs through interviews and focus groups Develop user personas
Development	Rapid and iterative prototyping of an AI model to maintain desired performance characteristics through testing with intended end-users.	Conduct iterative user testing Perform cognitive walkthrough Perform final (summative) usability testing
Implementation	Technical integration, testing and deployment, educational sessions for users, and consideration of interaction with other clinical systems, tools, and work processes.	Redesign existing workflows and processes to integrate new technology Conduct pilot test Refine based on user feedback
Long-term use	AI models should be continually monitored and validated to maintain desired performance and to detect safety events. Models may be retrained and additionally re-evaluated for modification of human factors elements.	Monitor user interaction data Provide a mechanism to report safety issues Monitor performance outcomes

Filice, R. W., & Ratwani, R. M. (2020). The Case for User-Centered Artificial Intelligence in Radiology