Lecture #2: Implementing AI in Healthcare (part 1)

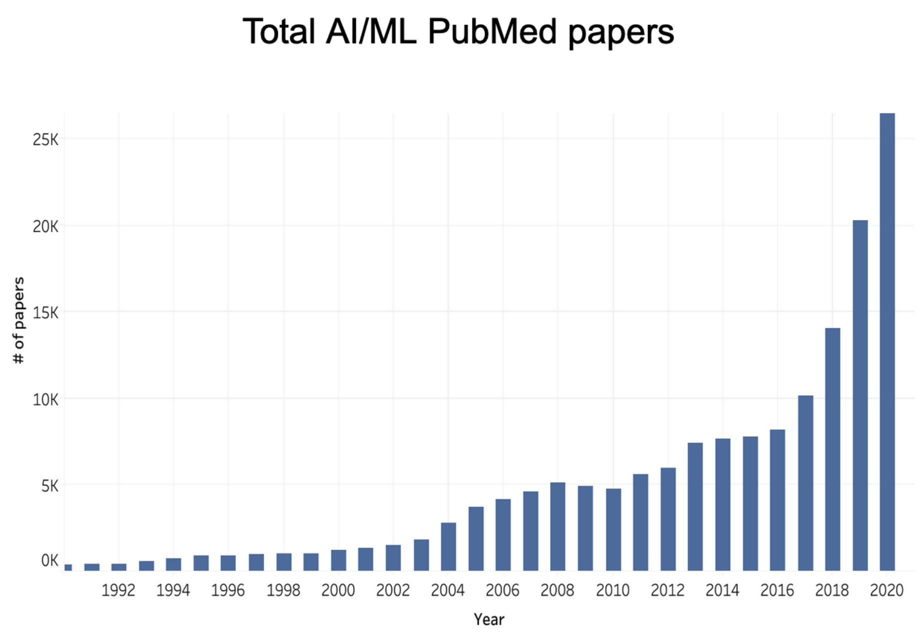

```
Data Sciences Institute
Topics in Deep Learning
Instructor: Erik Drysdale
TA: Jenny Du
```

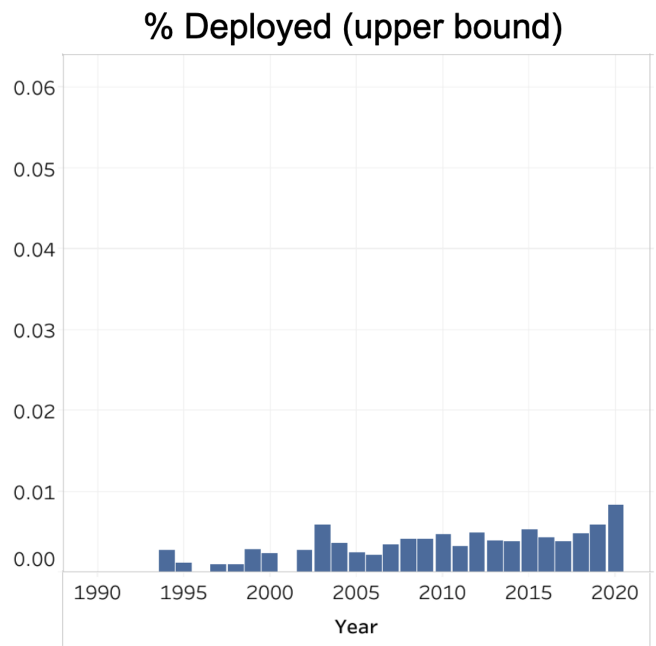
Lecture Outline

- The key difficulty (recap)
- The roadmap
- Choosing the right problem
- Developing a useful solution
- Considering ethical implications
- Rigorous evaluation (silent trial)
- Real-world examples
- Deploying AI/ML (MLOps)

Recap: technology adoption in HC is hard

We've got publications figured out





Source: MLHC 2021 (Anna Goldenberg)

Talk is cheap

How AI could change the future of our health care

CBC News: The National

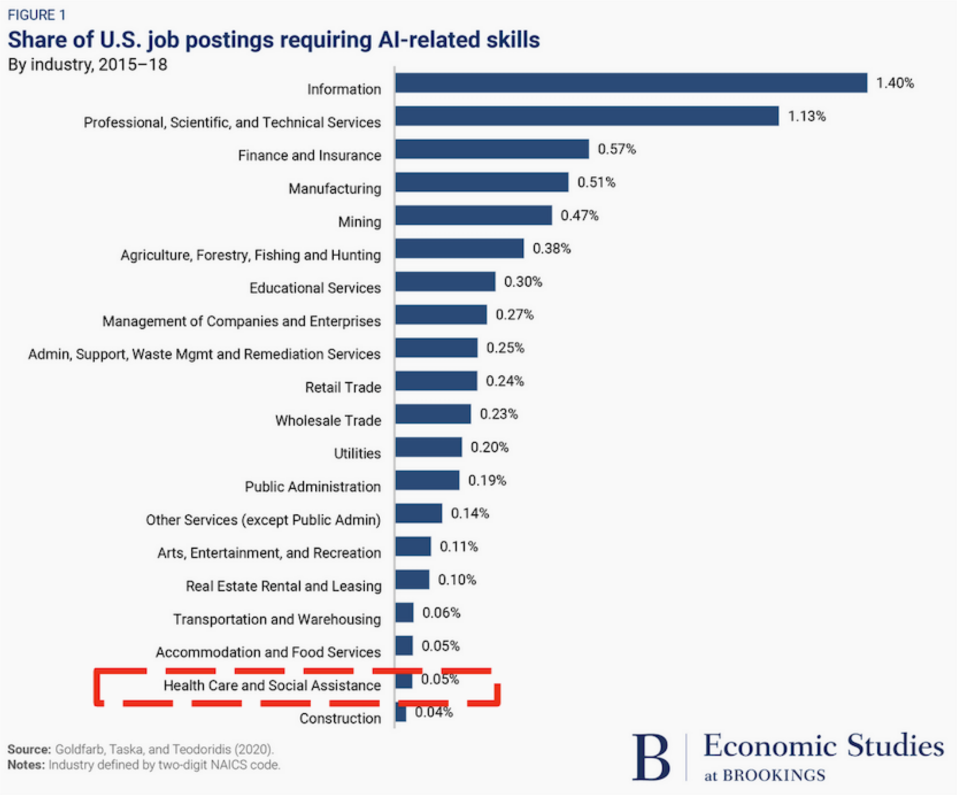
14K views • 4 years ago

RESEARCH

Why is AI adoption in health care lagging?

Avi Goldfarb, Florenta Teodoridis

March 9, 2022



Beware, beware ....

- **"The reality is that most failures of AI projects are failures in strategy and in execution."**
  - *Source: Why AI investments fail to deliver*
- **"Validation of the performance of an algorithm in terms of its accuracy is not equivalent to demonstrating clinical efficacy. This is ... the 'AI chasm'—that is, an algorithm with an AUC of 0.99 is not worth very much if it is not proven to improve clinical outcomes."**
  - *Source: High Performance Medicine*

Electronic Health Record (EHR) Data

- Challenges and barriers to achieving economies of scale in analyzing EHR data.
- Challenges in scaling EHR data analytics due to non-standardized systems.
- Importance of policy reforms and technology adoption for improved healthcare analysis and outcomes.

Where many fear to tread

future tense

How IBM's Watson Went From the Future of Health Care to Sold Off for Parts

BY LIZZIE O'LEARY JAN 31, 2022 • 9:00 AM

HEALTH TECH

Epic's widely used sepsis prediction model falls short among Michigan Medicine patients

By Dave Muoio • Jun 22, 2021 03:35pm

ARTIFICIAL INTELLIGENCE

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

NICOLE KOBIE BUSINESS 23.08.2022 12:00 PM

Babylon Disrupted the UK's Health System. Then It Left

A lot of upside

The Potential Impact of Artificial Intelligence on Healthcare Spending

Nikhil Sahni, George Stein, Rodney Zemmel & David M. Cutler

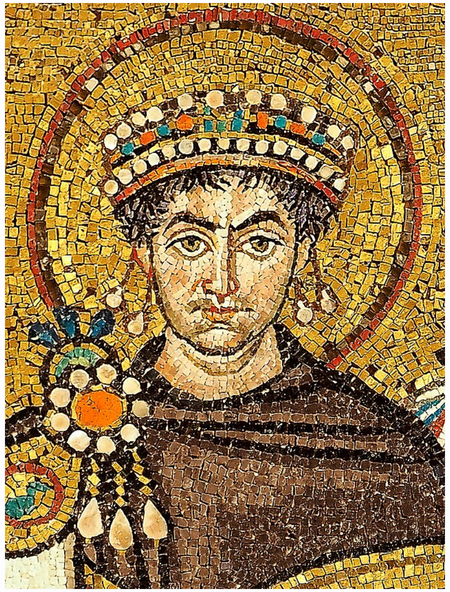
Figure 6. Breakdown of overall AI net savings opportunity within next five years using today's technology without sacrificing quality or access

Stakeholder group	Total costs (2019), $ billions	Net savings opportunity (2019), $ billions	Net savings opportunity as percent of stakeholder group's total costs	Percent of net savings opportunity focused on administrative costs
Hospitals	$1,096	$60–$120	5–11%	~40%
Physician groups	$711	$20–$60	3–8%	~50%
Private payers	$1,135	$80–$110	7–10%	~20%
Public payers	$511	$30–$40	5–7%	~20%
Other sites of care	$817	$10–$30	1–4%	~50%
Total		**$200–$360**	**5–10%[1]**	**~35%**

1. This represents the percent of total national health spending in 2019.

Source: National Health Expenditures data; authors' analysis

History's bumpy road of technological adoption



Justinian I

Emperor Hongxi

Sultan Bayezid II

Roadmap

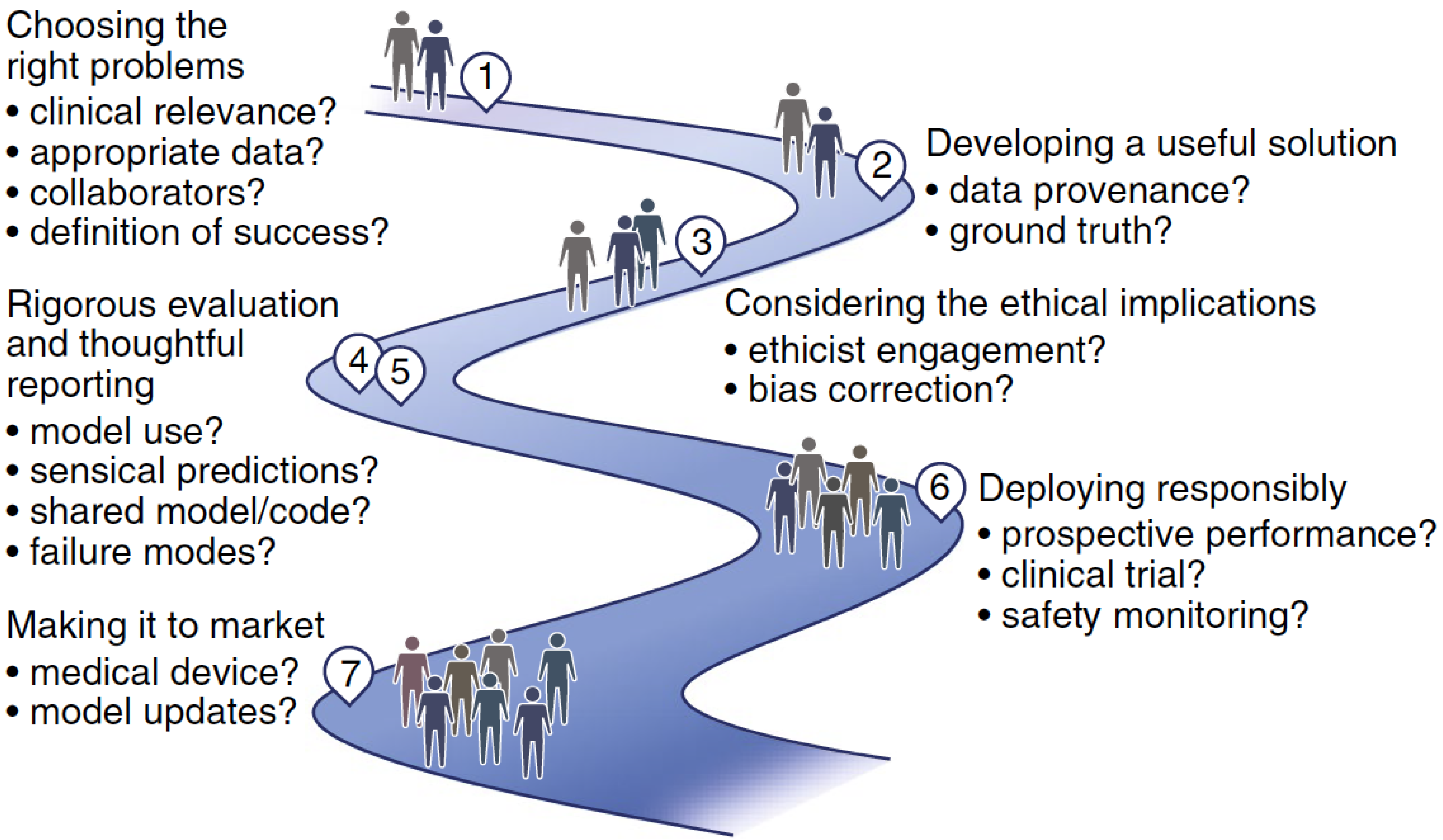
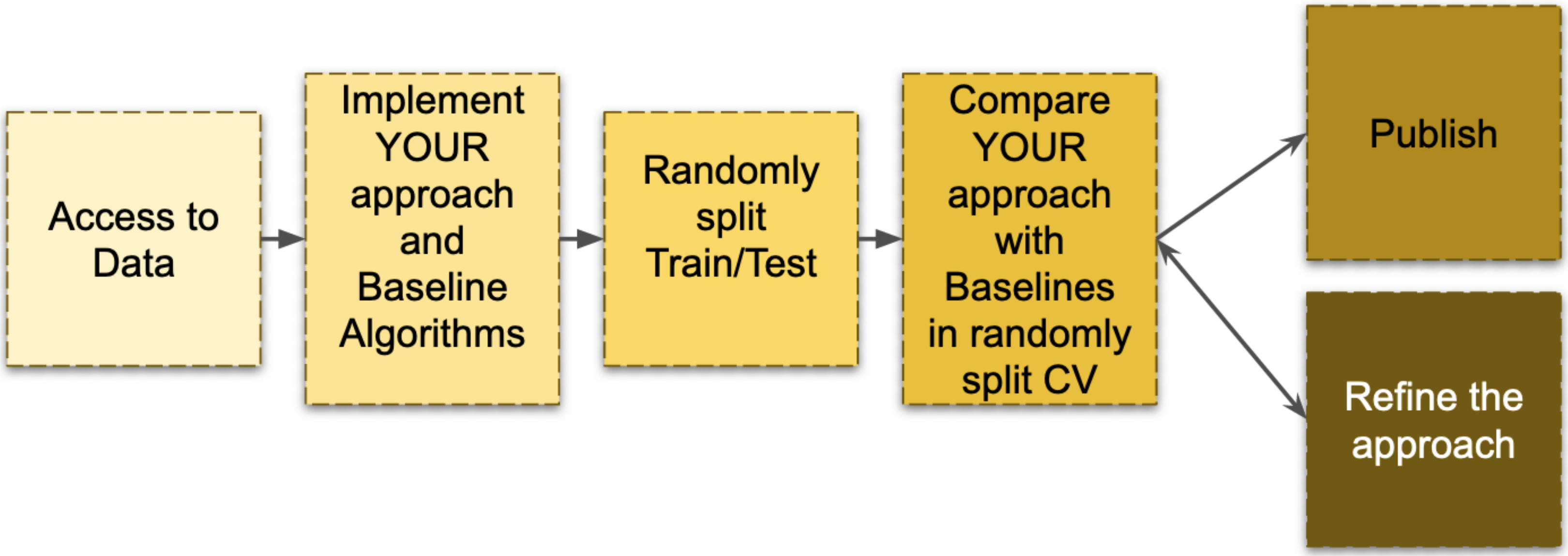
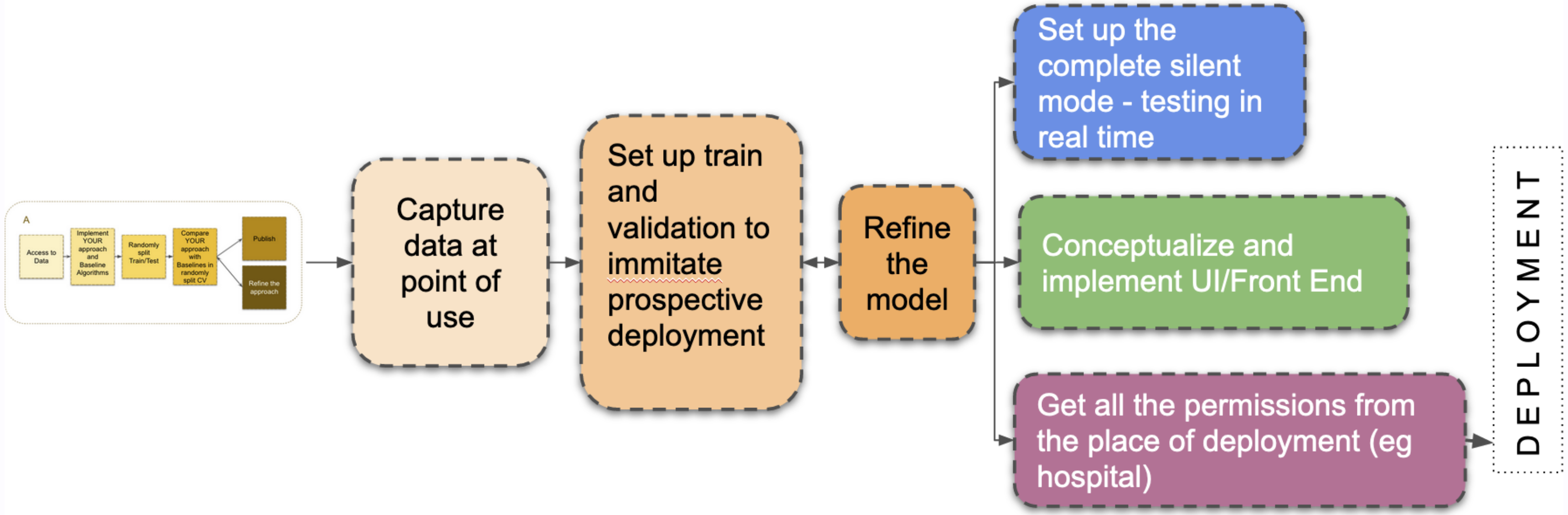


Fig. 1 | A roadmap for deploying effective ML systems in health care.
By following these steps and engaging relevant stakeholders early in the process, many issues stemming from the complexity of adopting ML in practice can be successfully avoided.

What we think ML scientists do



What ML scientists actually need to do



Choosing the right problem

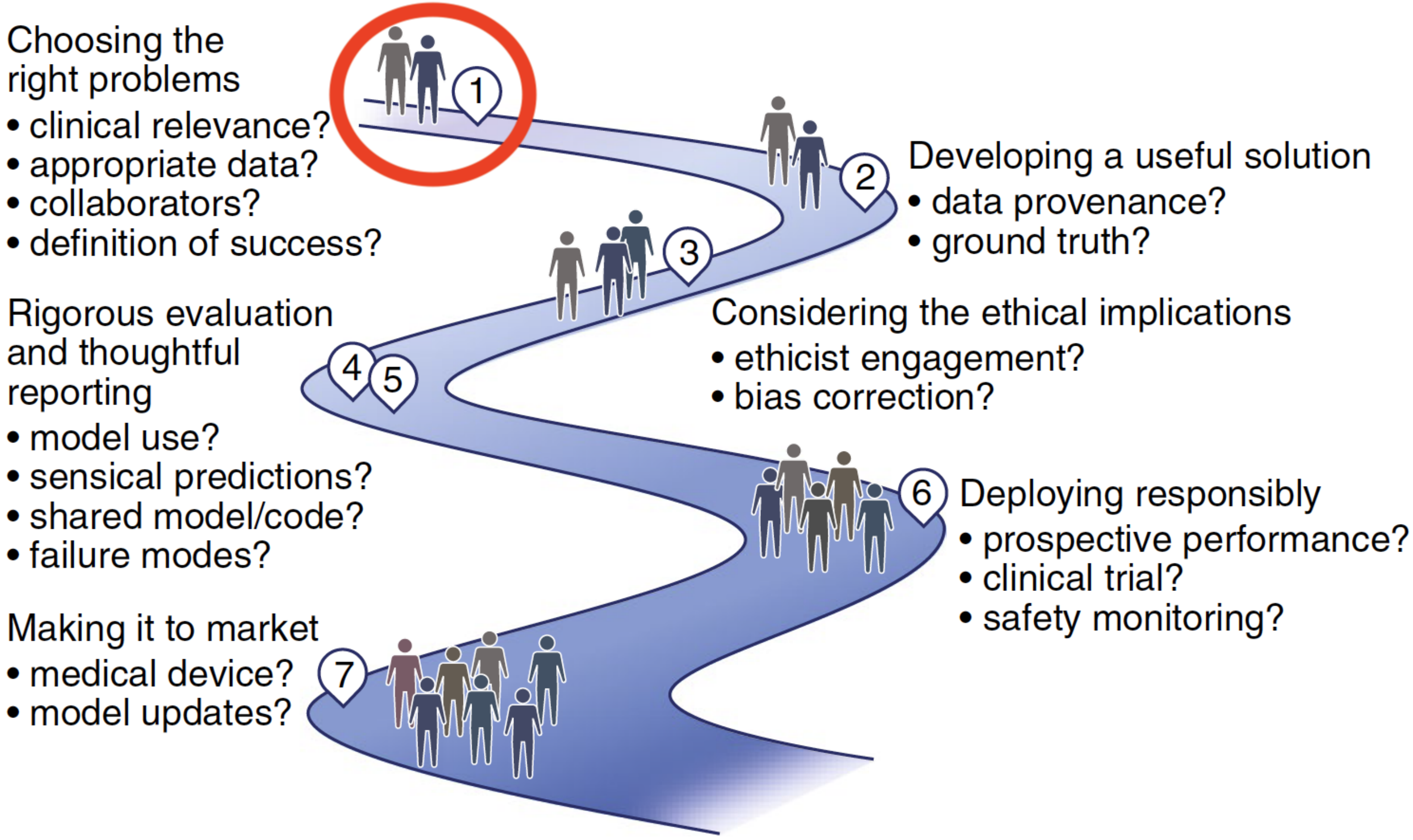
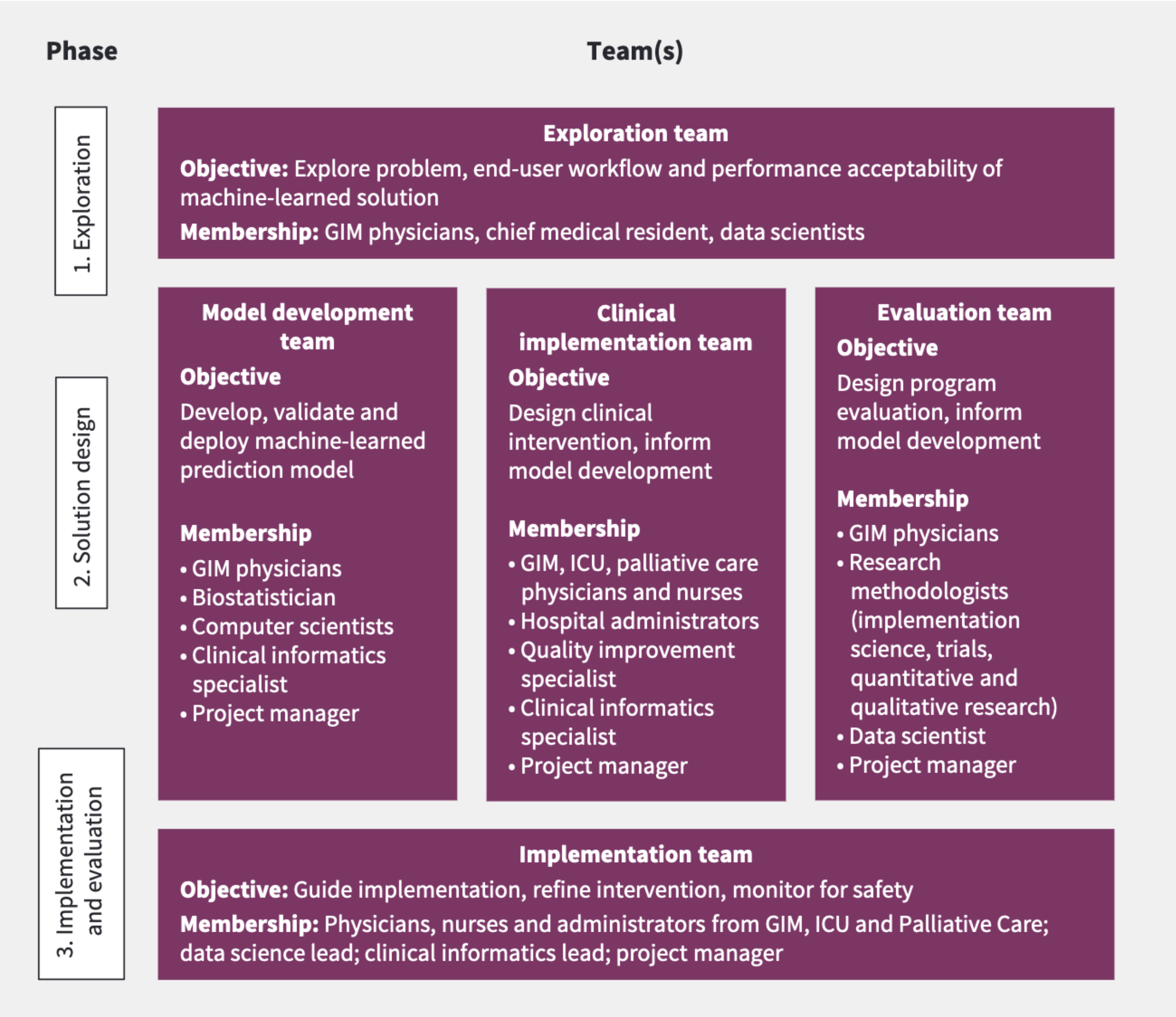


Fig. 1 | A roadmap for deploying effective ML systems in health care. By following these steps and engaging relevant stakeholders early in the process, many issues stemming from the complexity of adopting ML in practice can be successfully avoided.

Unclear problem formulations

- How does a prediction from a ML model actually help us solve a healthcare problem?
- How do we predict a "label" when there is no clear consensus among professionals?
- Who will help us navigate these problems?

Putting together the crew



Source: Verma et. al (2021)

Contextualization

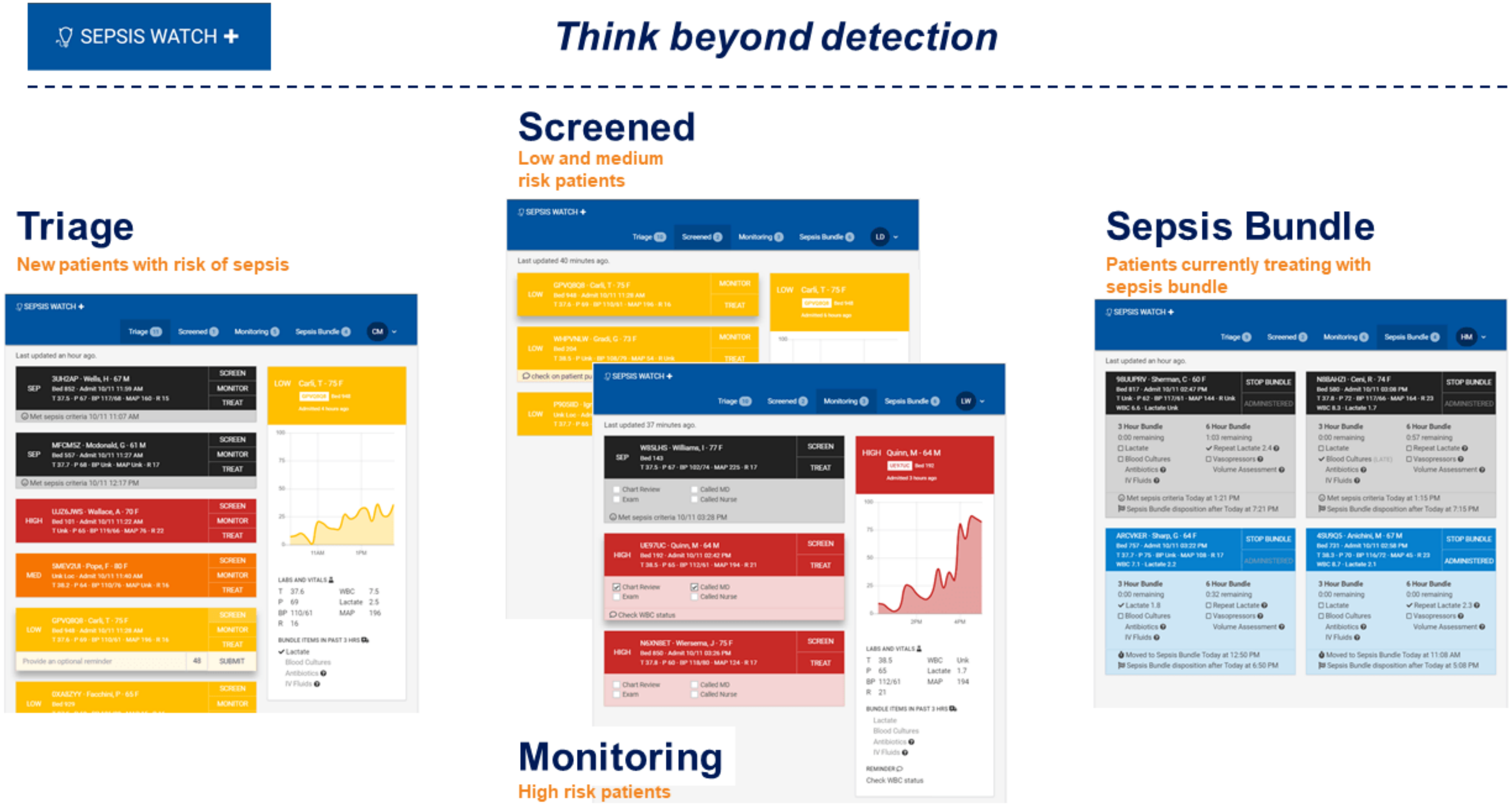
- AI/ML tools in healthcare must align with existing ways of working.
  - Understanding the current workflow is therefore crucial.
- For example, before developing an AI-based sepsis model you need to know:
  - How is Sepsis defined in the hospital? Does this vary between institutions
  - How is Sepsis currently detected? Which staff often spot it first? What point along the patient journey does this happen?
  - What actions are currently taken after a diagnosis?

The contextualization Mad Lib

- As a [decision maker],
- If I knew [information],
- I would do [intervention],
- To improve [measurable outcome]

If I knew [information]

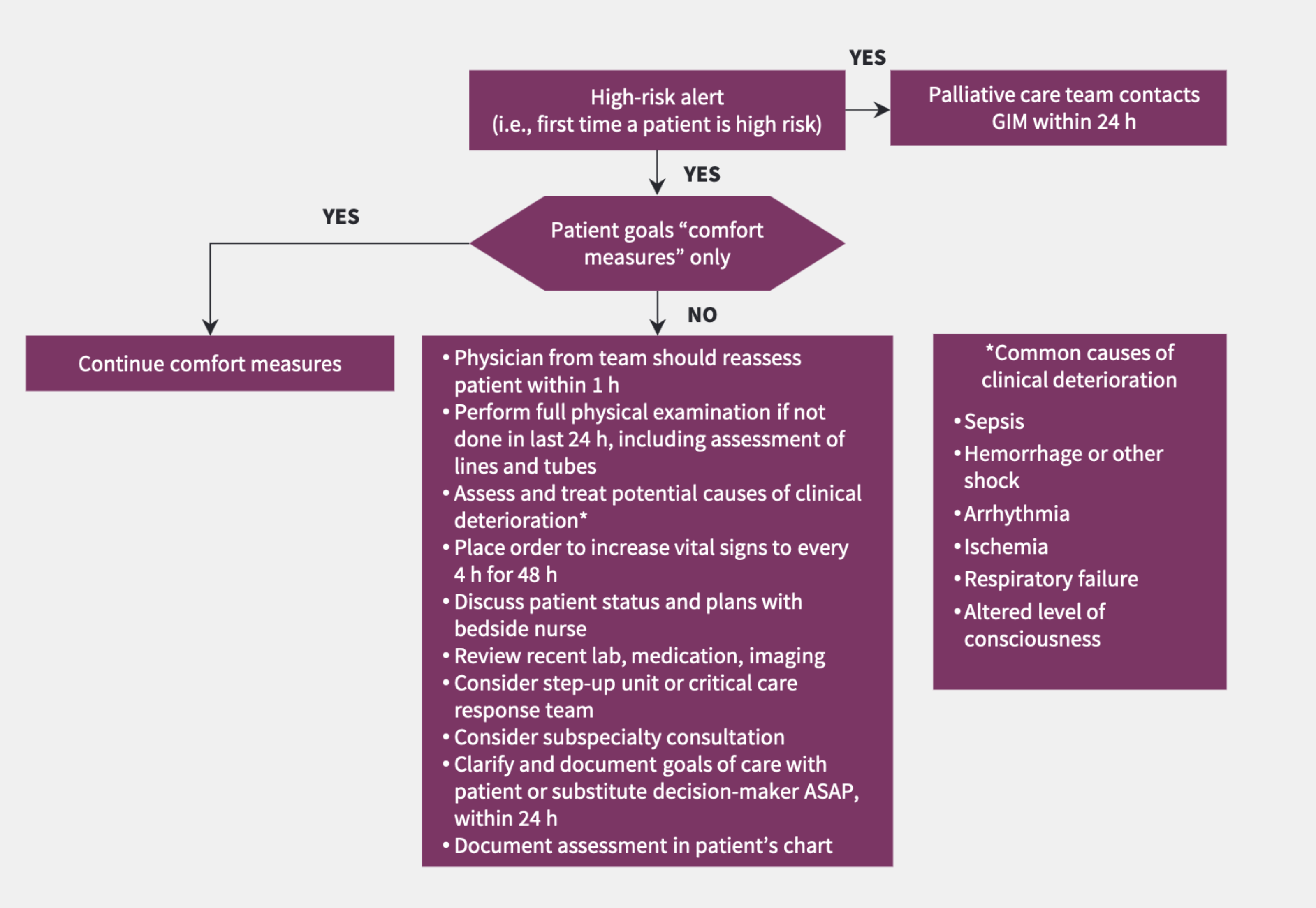
- A "middleman" can be used to parse informaiton (SepsisWatch)



Alert Fatigue

- Healthcare professionals receive an overwhelming number of alerts, which often leads to crucial warnings being ignored.
- Current approach to alerts does not consider human factors or user-centered design.
- Solutions include having a human vet which alerts (SepsisWatch), or calibrating a high precision classifier.

I would do [intervention]



Source: Verma et. al (2021)

Stakeholder engagement for problem definition

- Early stakeholder engagement identifies clinically relevant problems and ensures support throughout development.
  - *Note:* stakeholders can include healthcare providers, administrators, patients, and ethicists.
- Prioritizing clinically relevant and stakeholder-supported problems helps helps ensure diverse perspective and leads to impactful AI/ML solutions.
- **Rigorous problem definition** aligns solutions with stakeholder needs.

Stakeholder engagement: frontline health professionals

- Recognizing the expertise of frontline healthcare professionals is crucial.
- AI/ML tools should **augment clinical judgment, not replace it.**
- Involving clinicians in the development process is essential.
- Incorporating their feedback and providing necessary training and support enhances AI/ML effectiveness in practice.

Breakout #1

Suppose we wanted to reduce the rate of unplanned hospital re-admission (24H), what model would you build to do this, how would the model be acted on, and how would evaluate if it worked after deployment?

Developing a useful solution

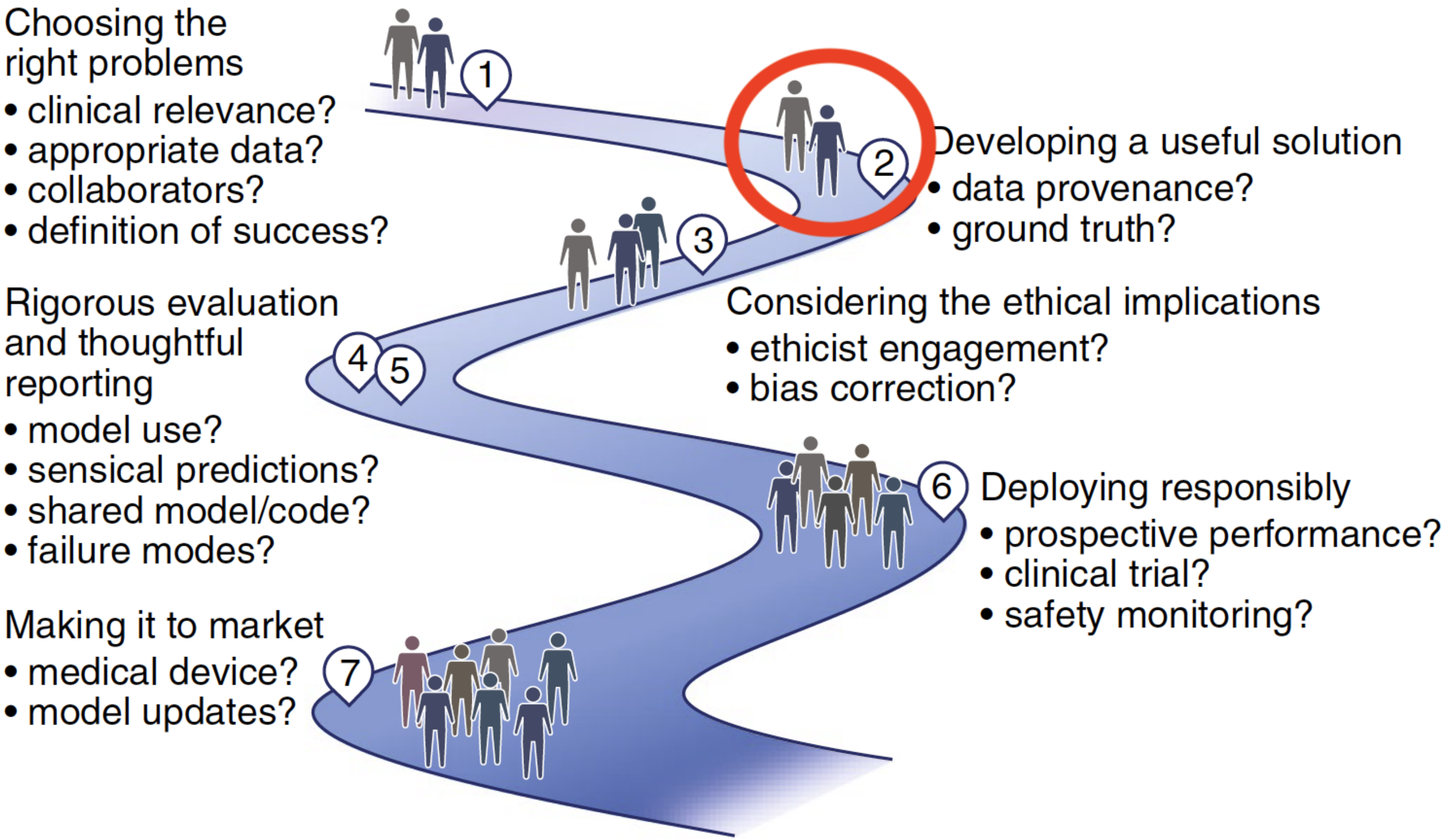
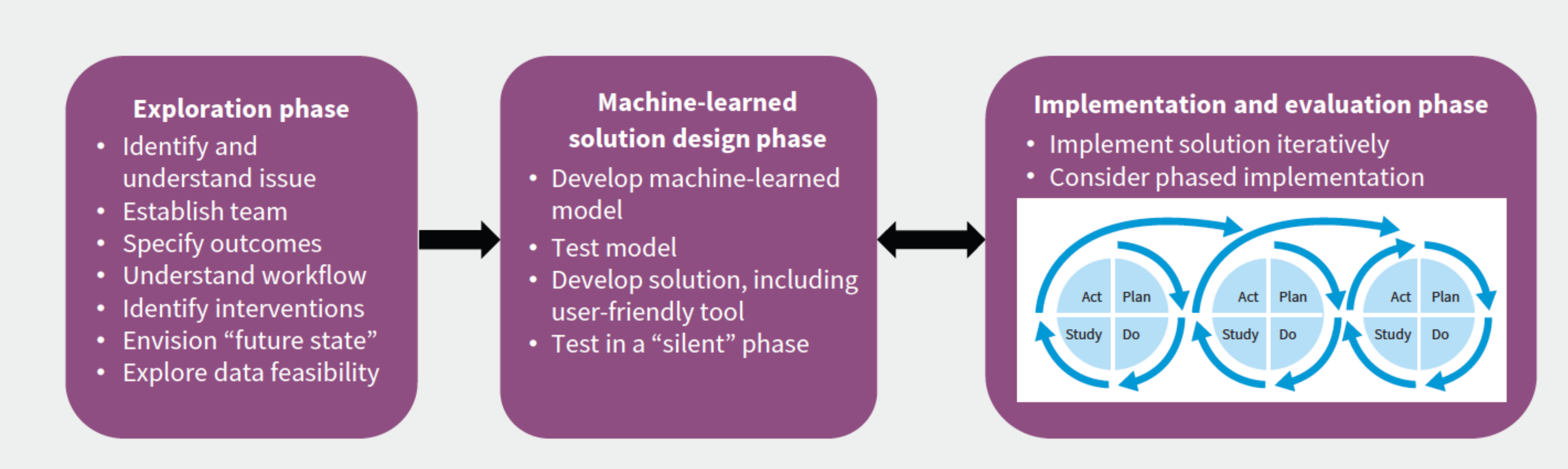


Fig. 1 | A roadmap for deploying effective ML systems in health care. By following these steps and engaging relevant stakeholders early in the process, many issues stemming from the complexity of adopting ML in practice can be successfully avoided.

Solution design

- AI/ML models and tools are developed based on the insights gained during the exploration phase.
- Emphasis is placed on designing solutions that are effective, interpretable, and usable by end-users.



Source: Verma et. al (2021)

Ground truth

- We take "labels" for granted in ML, we should't
  - Example of sepsis label shows complexity and opportunity for variability

2 or more SIRS criteria	- Temperature >38°C or <36°C (6 hours) - HR >90 (6 hours) - RR >20 (6 hours) - WBC count >12, <4, or % bandemia >10% (24 hours)
Suspect Infection	- Blood culture order (24 hours)
1 element of end organ failure	- Creatinine >2.0 (24 hours) - INR >1.5 (24 hours) - Total bilirubin >2.0 (24 hours) - SBP <90 or decrease in SBP by >40 (6 hours) - Platelets <100 (24 hours) - Lactate ≥2 (24 hours)

Data evaluation

- Before developing a solution, data must be thoroughly evaluated to ensure suitability for the problem at hand.
- Questions about data collection methods, purposes, and representativeness are crucial.
  - Ensure training data represent the environment where the model will be used.
  - Subtle biases in data can reduce model reliability and must be addressed during development.
  - Identifying and correcting biases upfront is crucial for model correctness.

Data quality control

- Conformance (fields aligns with expected format)
- Completeness (not missing)
- Plausability (believability, reasonability)

Analyte/Laboratory Measurement

Numeric	•Unit normalization performed through reference unit mapping •Parse specimen source value and subset by source •Parse typos and non-ASCII characters included in unit of measure •Apply project-specific or general upper and lower bound •Parse string text that conveys numeric result is outside range of measurement instrument and replace with upper bound •Ensure time stamp is captured in UTC	•Normalize serum creatinine values reported in mg/dL and mg/mL •Identifying and dropping serum glucose values that have specimen source of urine •Find and replace greek letters and typos used in units and replace with ASCII characters •Remove serum creatinine values over 150 mg/dL •Convert point-of-care glucose value of ">600 mg/dL" to "600 mg/dL" •Convert timestamp from EDT to UTC
Categorical	•Parsing specimen source value and subsetting by source •Map character string to hierarchical value through expert-derived reference table •Map character string to binary value through expert-derived reference table •Parse result text to identify indeterminate values •Ensure time stamp is captured in UTC	•Identify and drop blood cultures that have specimen source pleural fluid •Map blood culture results to negative, likely contaminant, and likely pathogen •Map HIV antibody titer levels to positive or negative result •Replace "hemolyzed sample" with missing value •Convert timestamp from EDT to UTC

Source: Sendak et. al (2022)

Considering ethical implications

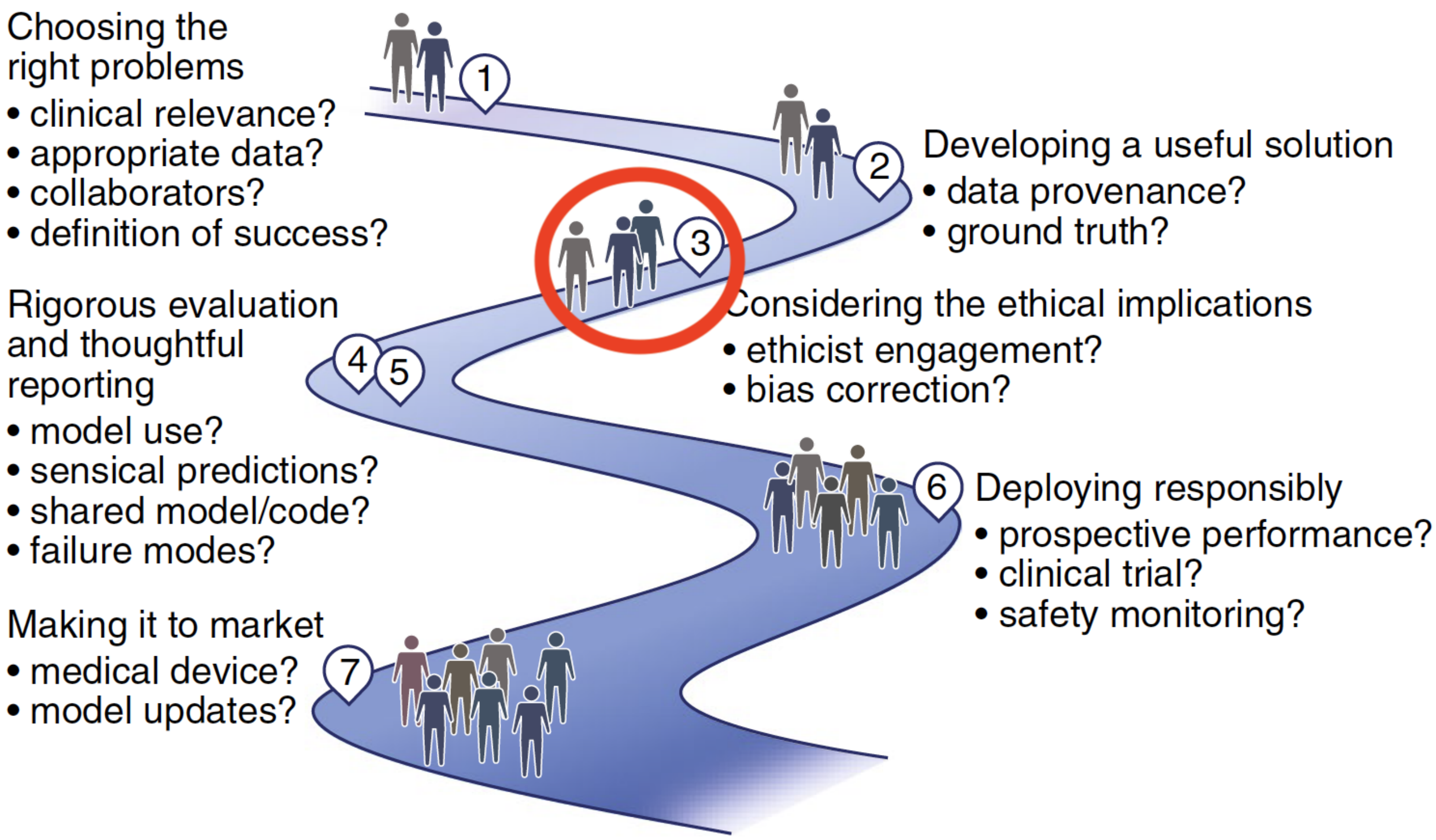


Fig. 1 | A roadmap for deploying effective ML systems in health care.
By following these steps and engaging relevant stakeholders early in the process, many issues stemming from the complexity of adopting ML in practice can be successfully avoided.

Health equity and disparities

- Health care data used for ML algorithms may be influenced by social inequalities (e.g., race, sex and other factors)
- Ethical questions may arise regarding the use of certain predictors, e.g., smoking status or HIV status
- Collaboration between ethicists, social scientists, regulatory scholars, AI/ML experts, and stakeholders is essential to address bias and ethical concerns.
- AI/ML algorithms focused on fairness can help mitigate biases and promote equitable healthcare delivery.

Rigorous evaluation and reporting

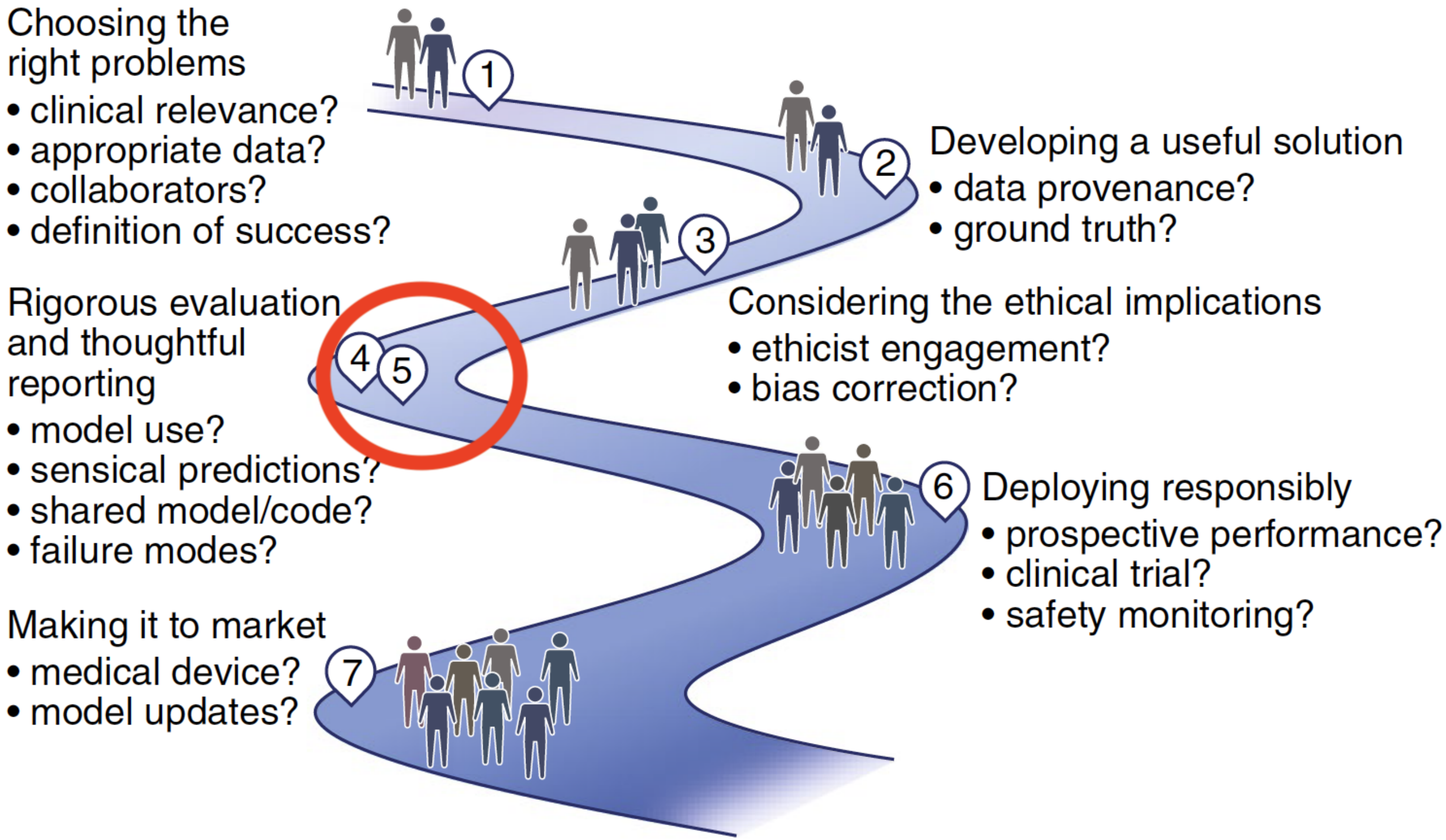


Fig. 1 | A roadmap for deploying effective ML systems in health care. By following these steps and engaging relevant stakeholders early in the process, many issues stemming from the complexity of adopting ML in practice can be successfully avoided.

Proper model evaluation

- Focus on **clinically relevant evaluation metrics**.
- Use qualitative approaches to uncover concerns missed by quantitative measures.
- Report results and share code and documentation for transparency.

Recall the Sepsis Model

- **Epic Sepsis Model Issues → Lack of reproducibility:**
  - Peer-reviewed data questioned the effectiveness of Epic's sepsis prediction algorithm.
  - University of Michigan Medical School study with over 27,000 patients found its performance "substantially worse" than reported.
- **Study Concerns:**
  - Lack of *external validation* for proprietary models and a call for transparency and validation before widespread clinical use.

Prospective evaluation

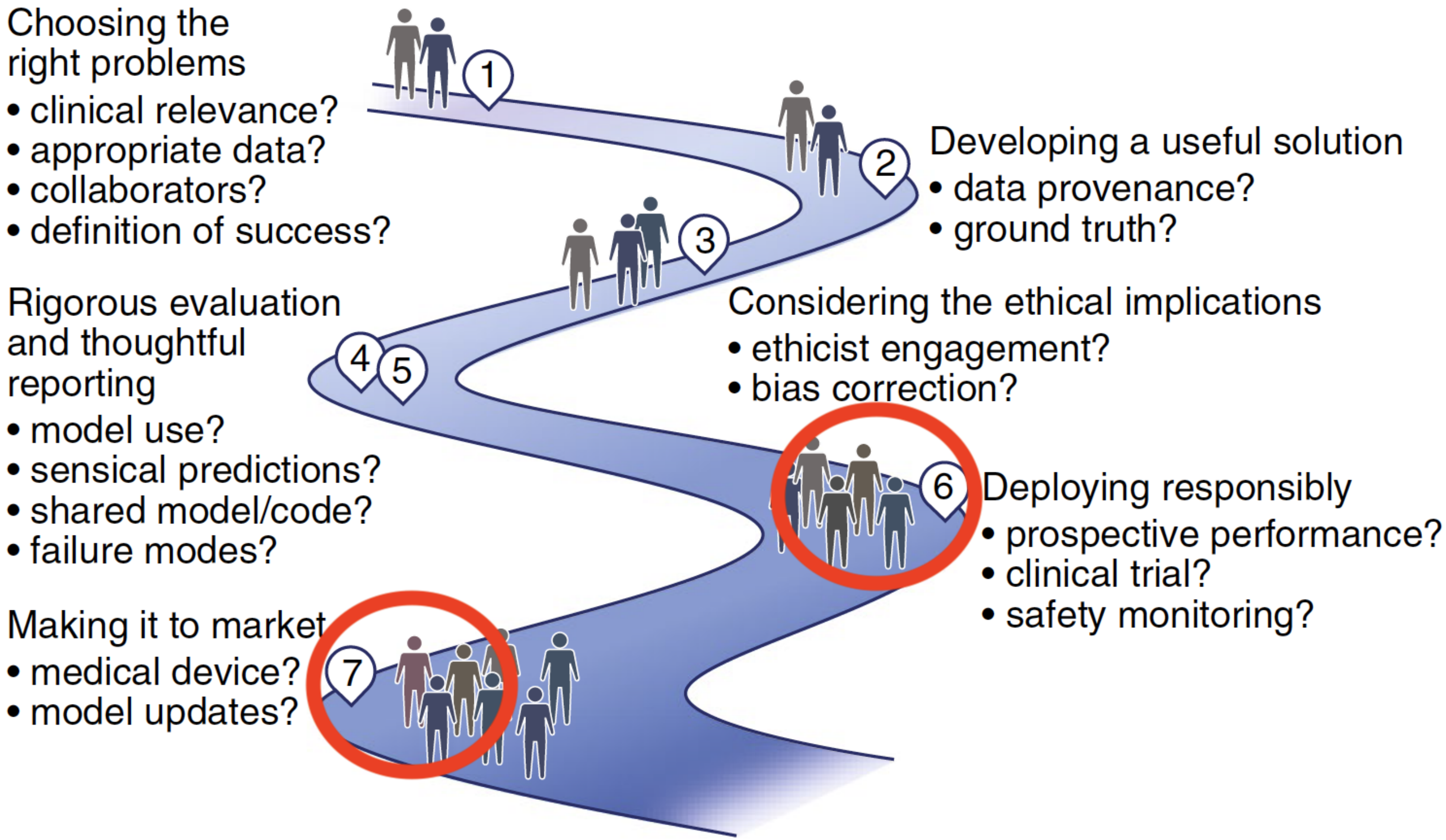
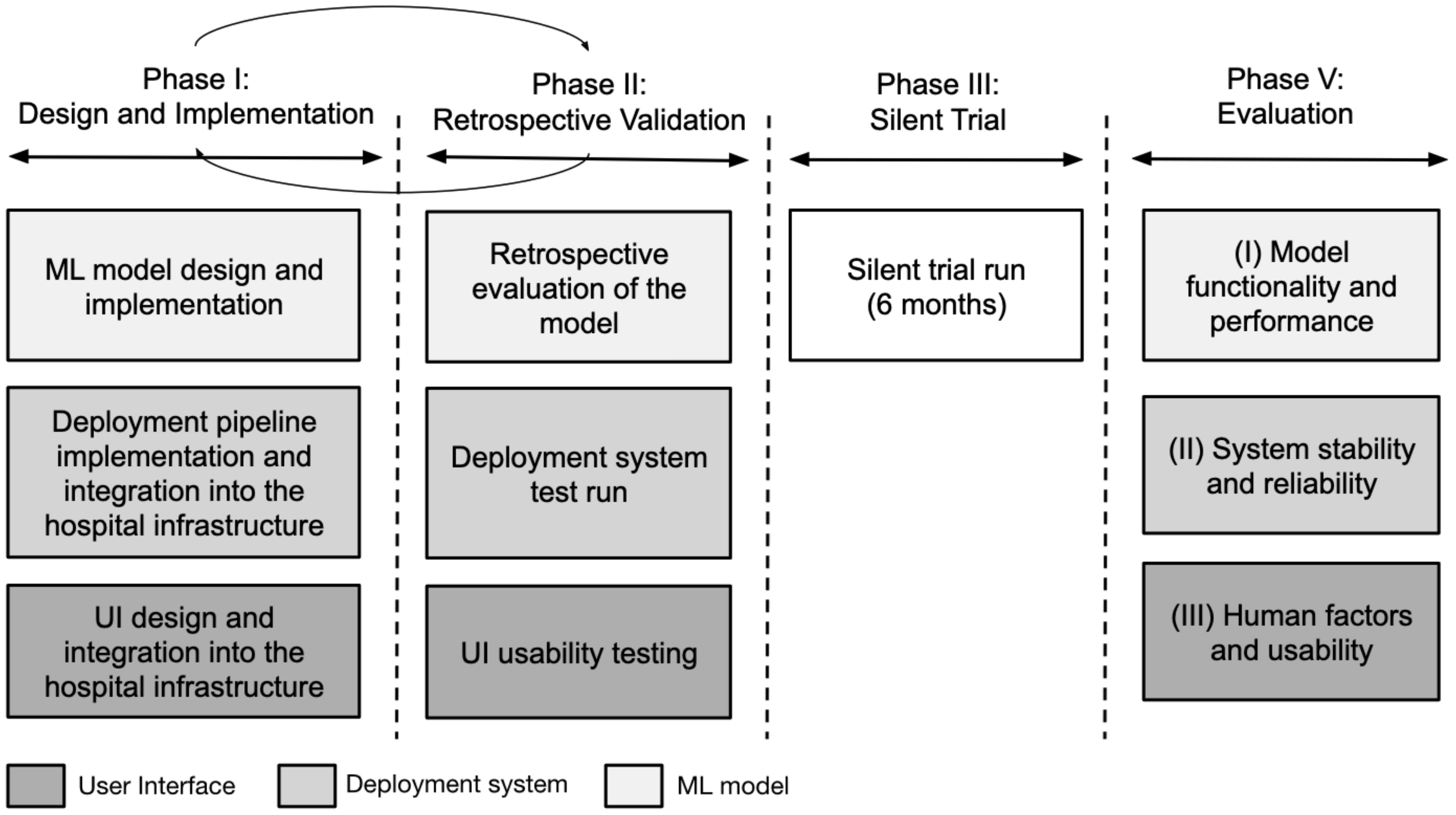


Fig. 1 | A roadmap for deploying effective ML systems in health care. By following these steps and engaging relevant stakeholders early in the process, many issues stemming from the complexity of adopting ML in practice can be successfully avoided.

Silent trial (overview)

- ML models **need** a real-time prospective evaluation to assess performance, failure points, and biases (w/ no human in-the-loop)



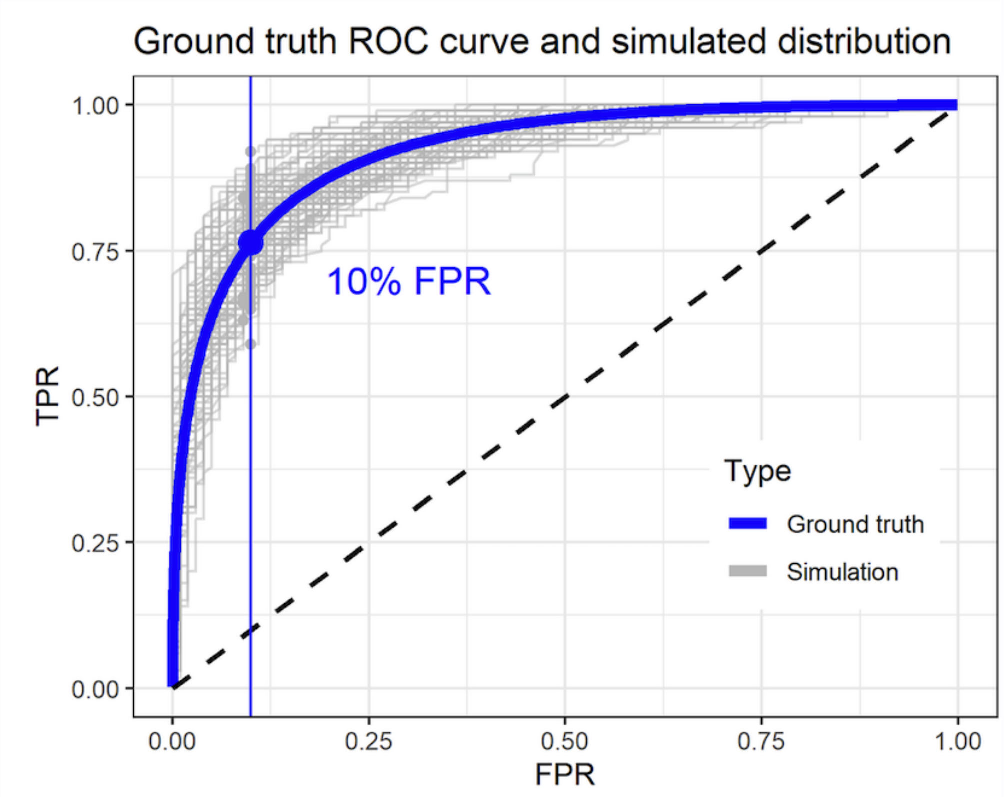
Source

Silent trial (statistics)

- A prospective trial let's us make one of two (statistical) claims rigorously:
  - This algorithm has at least [X] [accuracy measure]
  - This algorithm has at most [Y] [loss measure]
- How do we do this?
  - "Calibrate" model to have E[X] > X_hypothesis
  - ... or E[Y] < Y_hypothesis

Silent trial (statistics)

- In the case of a binary classifier, you need to pick an operating threshold to target a performance measure (e.g. sensitivity)
- But the operating performance=f(threshold) is a random variable

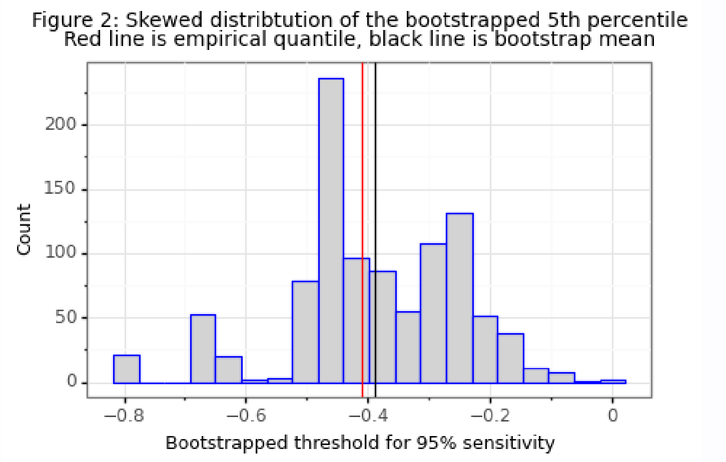


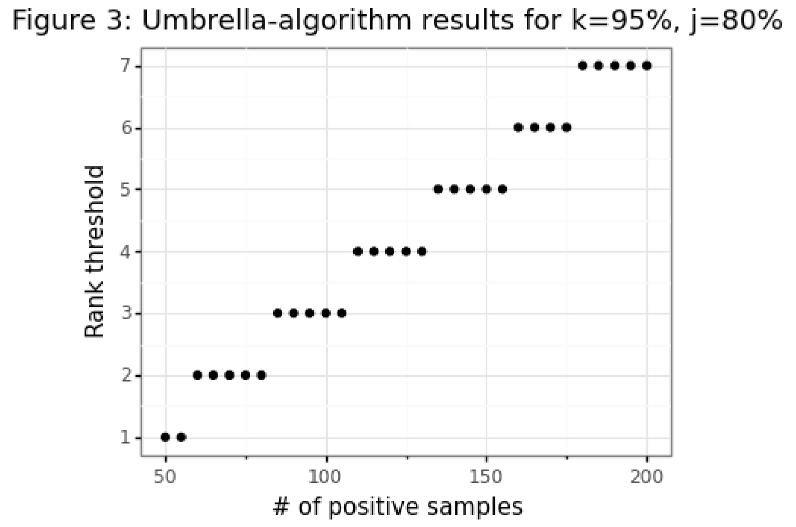
Breakout #2

Suppose we wanted to run a silent trial to demonstrate a model has at least 80% sensitivity, how would we "calibrate" the model so that the trial would likely be successful?

Silent trial (calibration)

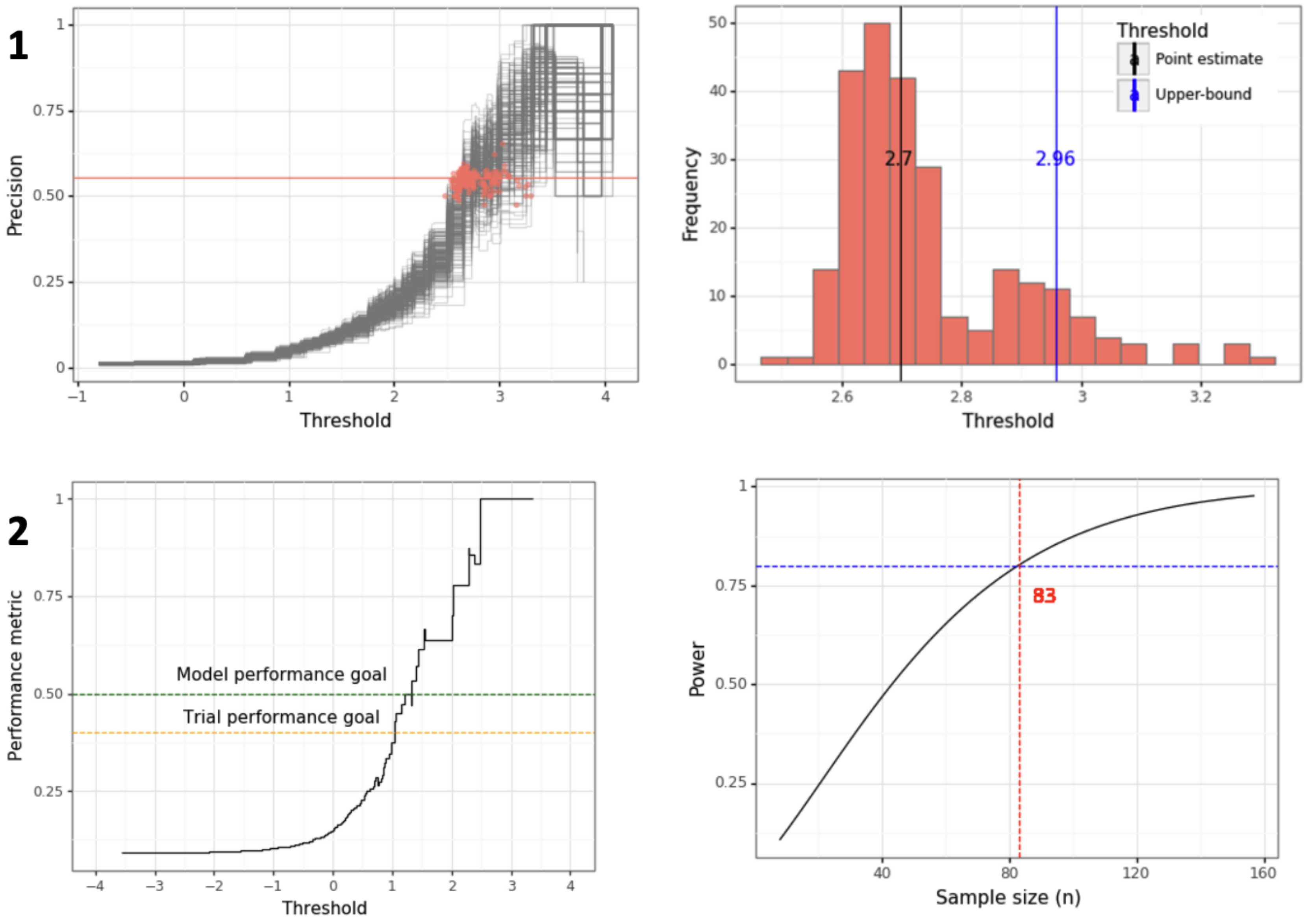
- You can increase the likelihood of success by picking a "conservative" operating threshold
  - Bootstrap or order statistic theory (see Tong et. al (2018))





Empirical bootstrap Rank order approach

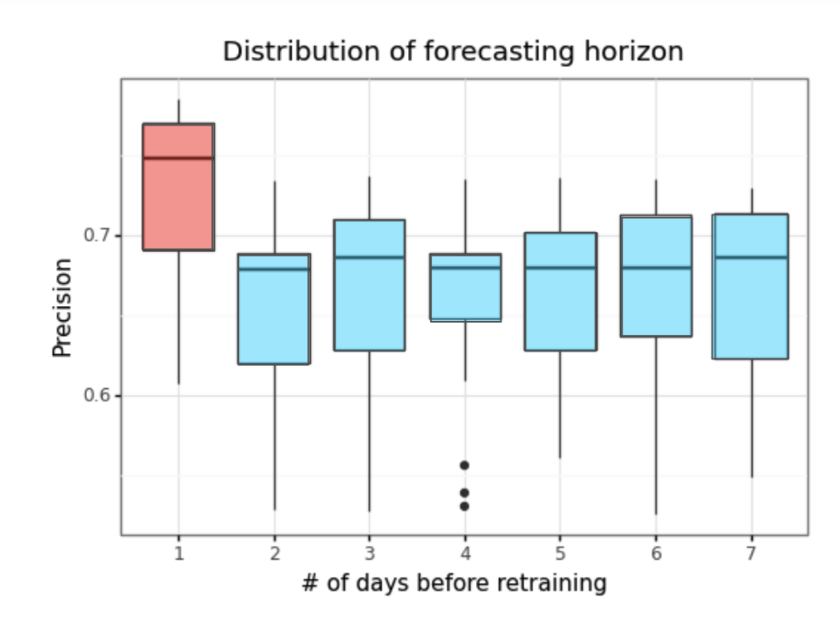
Statistical calibration (putting it together)



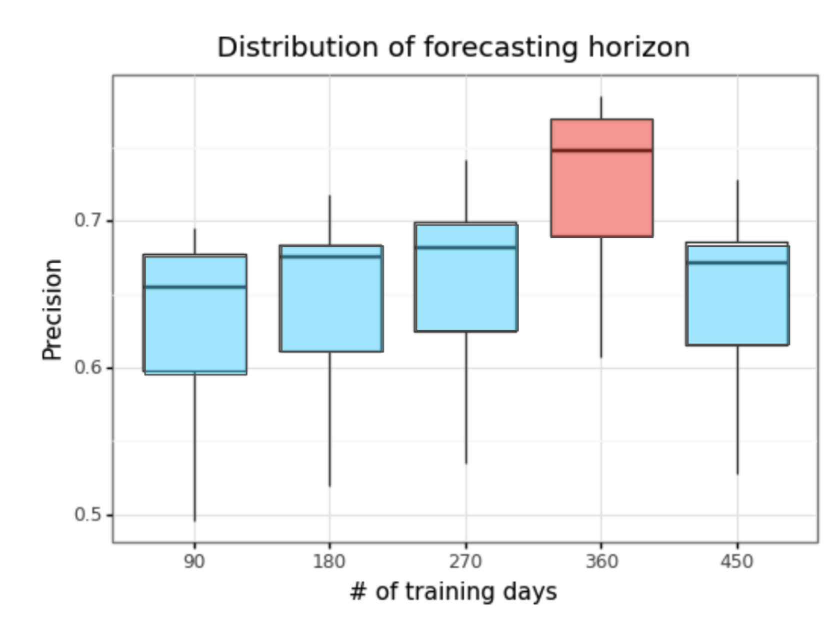
MLOps

Common deployment considerations

- How often to re-train, and with what data? Consider this an empirical question.



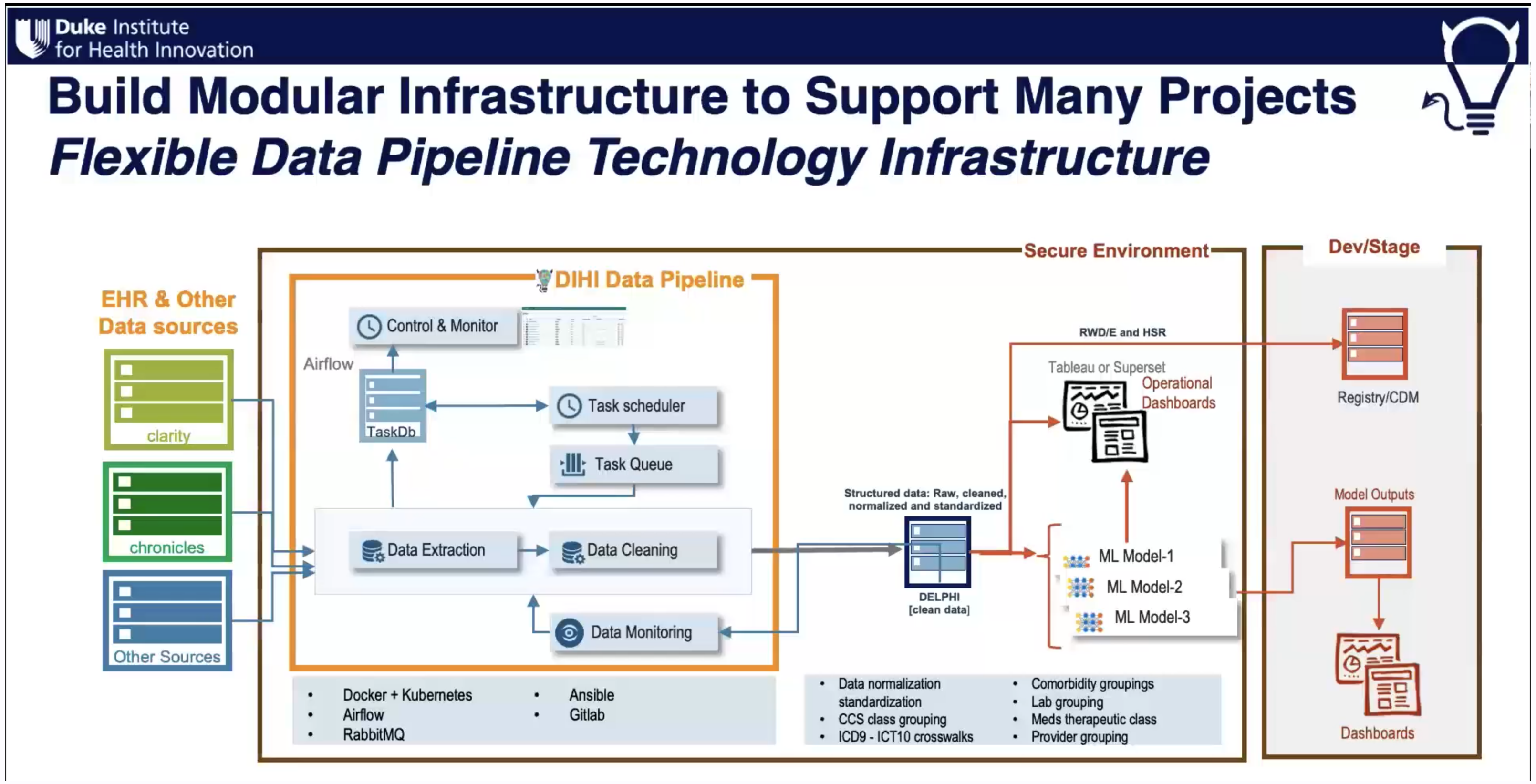
Frequency



Training window

Common deployment considerations

- Example from DIHI: Advancing the Safe, Effective and Equitable Use of AI in Healthcare



Implementation and evaluation

- **Continuous monitoring and feedback** mechanisms allow for iterative improvements to the tool over time.
- Ongoing evaluation helps identify and address any unintended consequences or disparities in healthcare deliver

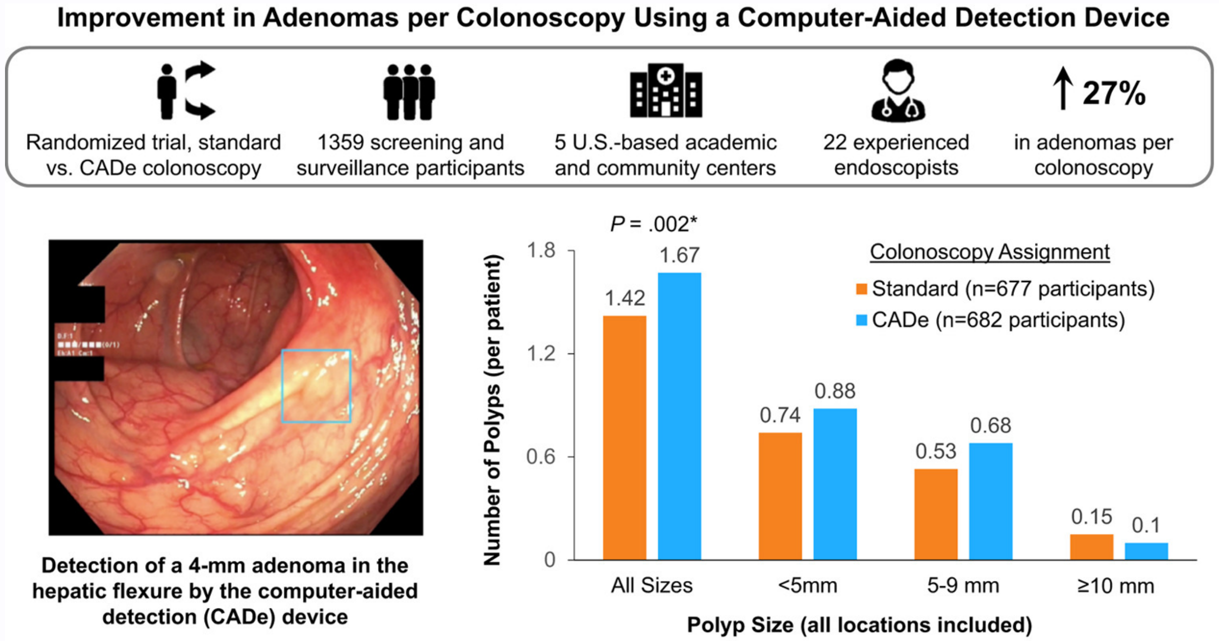
Summary

Considerations for successful translation of AI/ML into healthcare

- **Clear problem definition** is crucial for effective AI/ML deployment in healthcare.
- **Engaging stakeholders** early and into all stages of development ensures identification of clinically relevant problems.
- **Thorough data evaluation** is necessary to address biases and ensure alignment with existing workflows.
- **Continuous monitoring and feedback** in real-world settings are essential for successful AI/ML deployment.

Case Studies

Successes with colonoscopy



Gastroenterology

Source: Shauket et. al (2022)

Google's Diabetic Retinopathy

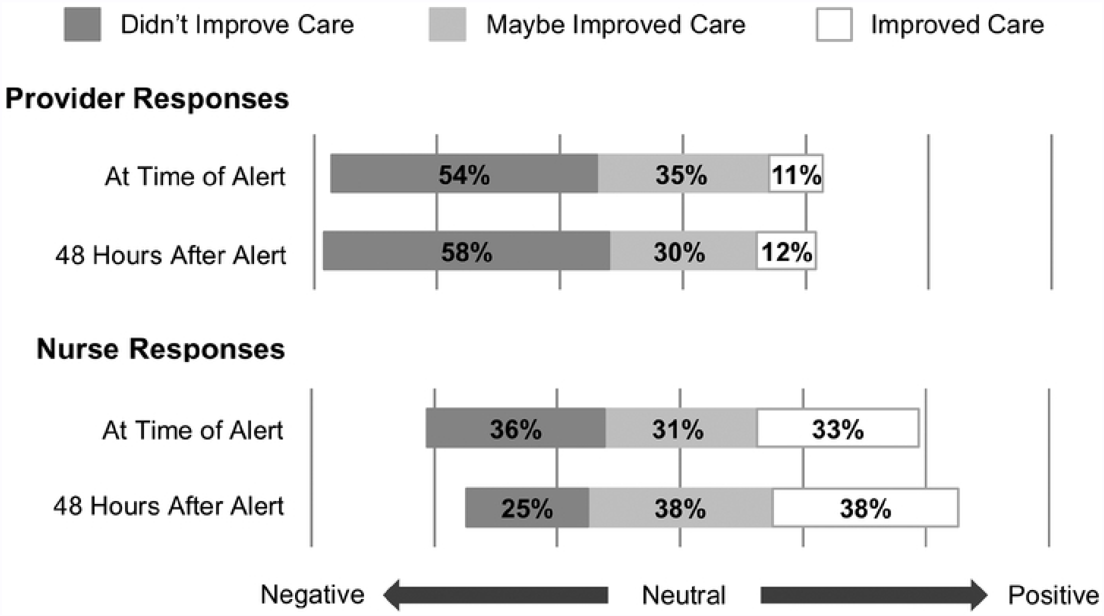
- Deployment of a deep learning system for diabetic retinopathy screening in Thai clinics (source).
- Key findings reveal the challenges of integrating AI into clinical workflows, such as issues with system gradability, internet connectivity, and the necessity of obtaining patient consent.

Nurse operates the takes images of patient’s retina

UPenn's Sepsis Model

- Developed and evaluated a machine learning algorithm aimed at predicting severe sepsis and septic shock within a tertiary teaching hospital system.
- Algorithm, based on a random-forest classifier and electronic health record data, showed a sensitivity of 26% and specificity of 98%.

- Implementation led to a modest increase in lactate testing and IV fluid administration but no significant change in mortality or ICU transfer rates, though it did reduce the time-to-ICU transfer.



Clinician perceived impact on patient care (https://doi.org/10.1097/ccm.0000000000003803).

UMichigan case management (similar story)

Is the model effective when used?

Karandeep Singh

The model was good.

AMIA Annu Symp Proc. 2018; 2018: 295–304.
Published online 2018 Dec 5.

PMCID: PMC6371247
PMID: 30815068

Towards a Learning Health System to Reduce Emergency Department Visits at a Population Level

Elliott Brannon, MPH,[1] Tianshi Wang,[2] Jeremy Lapedis, DrPH,, MHPS,[3] Paul Valenstein, MD,[4] Michael Klinkman, MD,, MS,[5] Ellen Bunting, MA,[6] Alice Stanulis,[6] and Karandeep Singh, MD,, MMSc[1,2,7]

▸ Author information ▸ Copyright and License information Disclaimer

Abstract Go to:

High utilizers of the Emergency Department (ED) often have complex needs that require coordination of care between multiple organizations. We describe a Learning Health Systems (LHS) approach to reducing ED visits, in which an intervention is delivered to a cohort of high utilizers identified using population-level data and predictive modeling. We focus on the development and validation of a random forest model that utilizes electronic health record data from three health systems across two counties in Michigan to predict the number of ED visits each resident will incur in the next six months. Using 5-fold cross-validation, the model achieves a root-mean-squared-error of 0.51 visits and a mean absolute error of 0.24 visits. Using time-based validation, the model achieves a root-mean-squared error of 0.74 visits and a mean absolute error of 0.29 visits. Patients projected to have high ED utilization are being enrolled in a community-wide care coordination intervention using twelve sites across two counties. We believe that the repeated cycles of modeling and intervention demonstrate an LHS in action.

But using it was ineffective.

Original Research | Published: 10 March 2021

Predictive Model-Driven Hotspotting to Decrease Emergency Department Visits: a Randomized Controlled Trial

Brady Post PhD, Jeremy Lapedis DrPH, Karandeep Singh MD, Paul Valenstein MD, Ayşe G. Büyüktür PhD, Karin Teske MPH & Andrew M. Ryan PhD

Journal of General Internal Medicine (2021) | Cite this article

Conclusions

The community case management intervention targeting ED visits was not associated with reduced utilization. Future case management interventions may benefit from additional patient engagement strategies and longer evaluation time periods.

Trial Registration

Clinicaltrials.gov Identifier: NCT03293160.

Figure adapted from:

> Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2022). Do No Harm: A Roadmap for Responsible Machine Learning in Healthcare. Nature Medicine

Ideas adapted from:

> Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2022). Do No Harm: A Roadmap for Responsible Machine Learning in Healthcare. Nature Medicine
>
> Drysdale, E., Dolatabadi, E., Chivers, C., Liu, V., Saria, S., Sendak, M., Wiens, J., Brudno, M., Hoyt, A., Mazwi, M., Mamdani, M., Singh, D., Allen, V., McGregor, C., Ross, H., Szeto, A., Anand, A., Verma, A., Wang, B., Paprica, P. A., & Goldenberg, A. (2020). Implementing AI in healthcare. Vector-SickKids Health AI Deployment Symposium, Toronto, Ontario, Canada.

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S., & O'Brien, C. (2020). "The human body is a black box": supporting clinical decision-making with deep learning. FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency

Verma, A. A., Murray, J., Greiner, R., Cohen, J. P., Shojania, K. G., Ghassemi, M., Straus, S. E., Pou-Prom, C., & Mamdani, M. (2021). Implementing machine learning in medicine. CMAJ.