Institution-Wide Governance for AI in Healthcare

Anand Chowdhury, MD, MMCi

Annual Conference 2024

Building the Future of Health Together

Presenter

Anand Chowdhury, MD, MMCi

Director of Informatics for Artificial Intelligence, Duke Health

Agenda

- Discuss the importance of Institutional Governance for AI
- Review the ABCDS Process at Duke
- Adaptations made for Generative AI
- Use cases

Learning Objectives

- Describe the opportunities and challenges of using generative AI in clinical decision support.
- Apply the ABCDS Oversight Framework for the governance and evaluation of generative AI tools.
- Summarize the best practices and guidelines established for the responsible use of generative AI in healthcare.

Promise of Artificial Intelligence/Machine Learning in Health Care

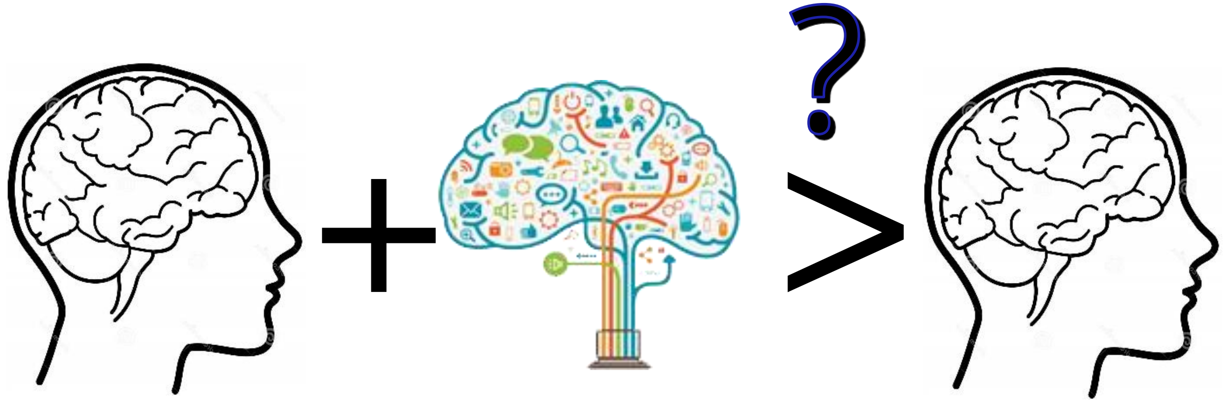
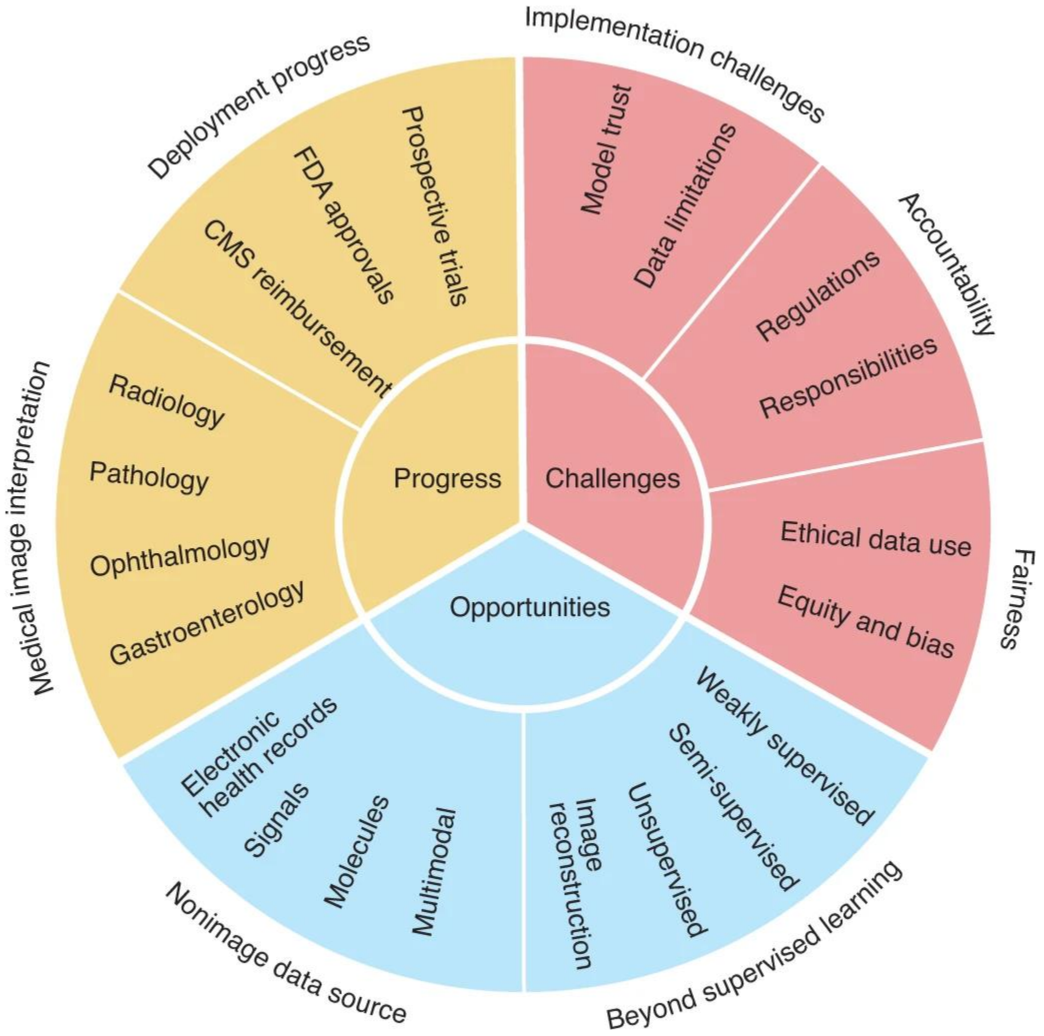


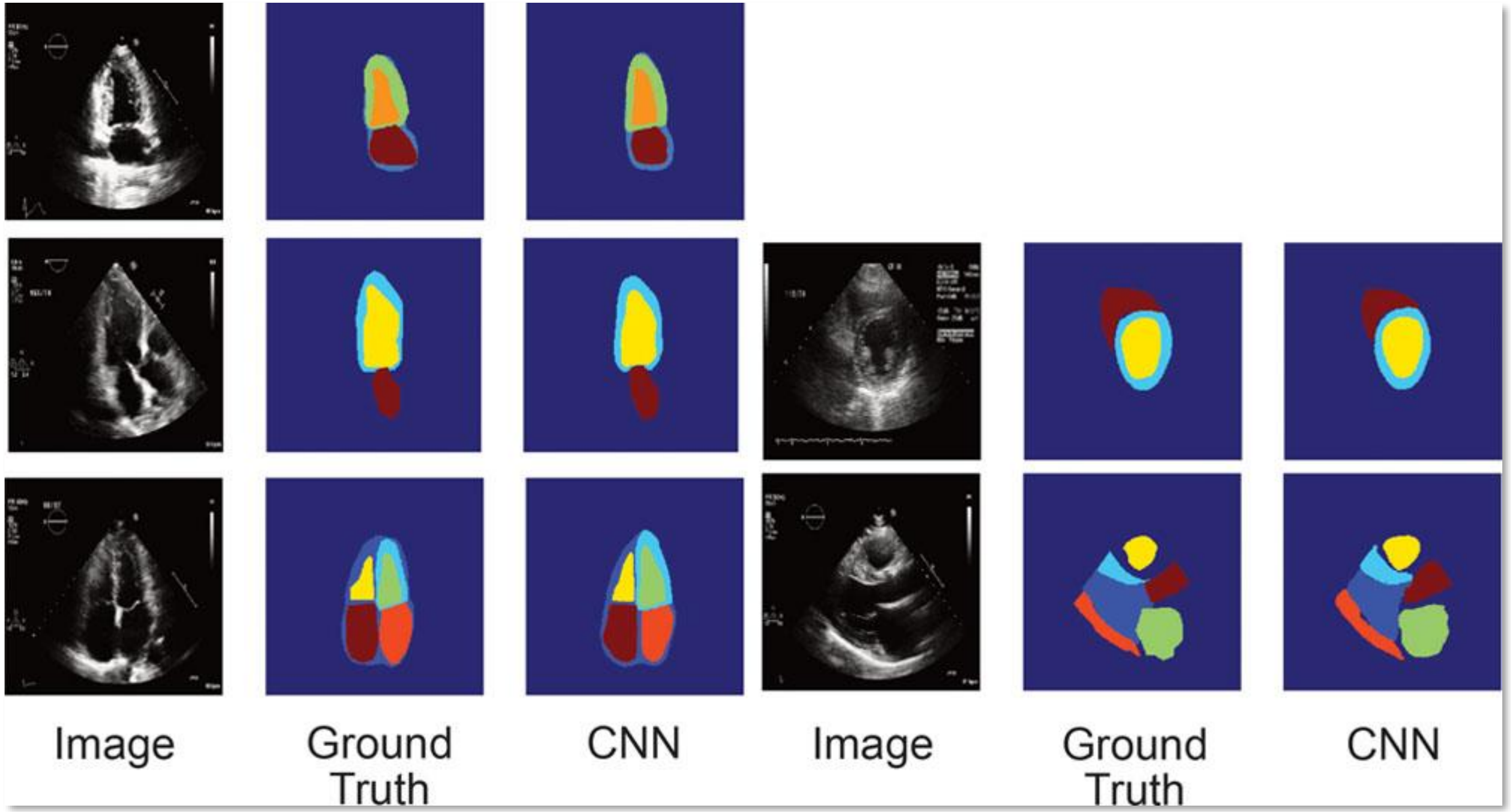
Photo by John McArthur on Unsplash

The sky is the limit



Nat Med. 2022;28(1):31-38.

Computer Vision for Cardiac Ultrasound



Circulation. 2018;138(16):1623-1635.

Population Health

REVIEW ARTICLE OPEN

Check for updates

Big data, machine learning, and population health: predicting cognitive outcomes in childhood

Andrea K. Bowe[1], Gordon Lightbody[1,2], Anthony Staines[3] and Deirdre M. Murray[1]

Pediatr Res. 2023;93(2):300-307

Predicting Preclinical Heart Failure Progression

The Rise of Machine-Learning for Population Health*

Jordan B. Strom, MD, MSc,[a,b,c] Partho P. Sengupta, MD, DM[d]

JACC Cardiovasc Imaging. 2022;15(2):209-211.

“Wild West” of Algorithms

“We have a Wild West of algorithms,” said Michael Pencina, coalition [CHAI] co-founder and director of Duke AI Health. There’s so much focus on development and technological progress and not enough attention to its value, quality, ethical principles or health equity implications.”

Politico, April 4, 2023

AI/ML Risks

CHEST JOURNAL

Log in

Search for...

EDUCATION AND CLINICAL PRACTICE: GUIDELINES AND CONSENSUS STATEMENTS | VOLUME 164, ISSUE 2, P461-475,

*“**Several assumptions and gaps**, both in the published literature and in our evolving understanding of lung health, were identified. It seems that many past perceptions and practices regarding the effect of race and ethnicity on PFT results interpretation are based on **limited scientific evidence and measures that lack reliability**.”*

RESEARCH

Science AAAS

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich

*“At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. **Remedying this disparity** would increase the percentage of Black patients receiving additional help from **17.7%** to **46.5%**. The bias arises because the algorithm predicts health care costs rather than illness…”*

Science. 2019;366(6464):447-453.
CHEST. 2023;164(2):461-475.

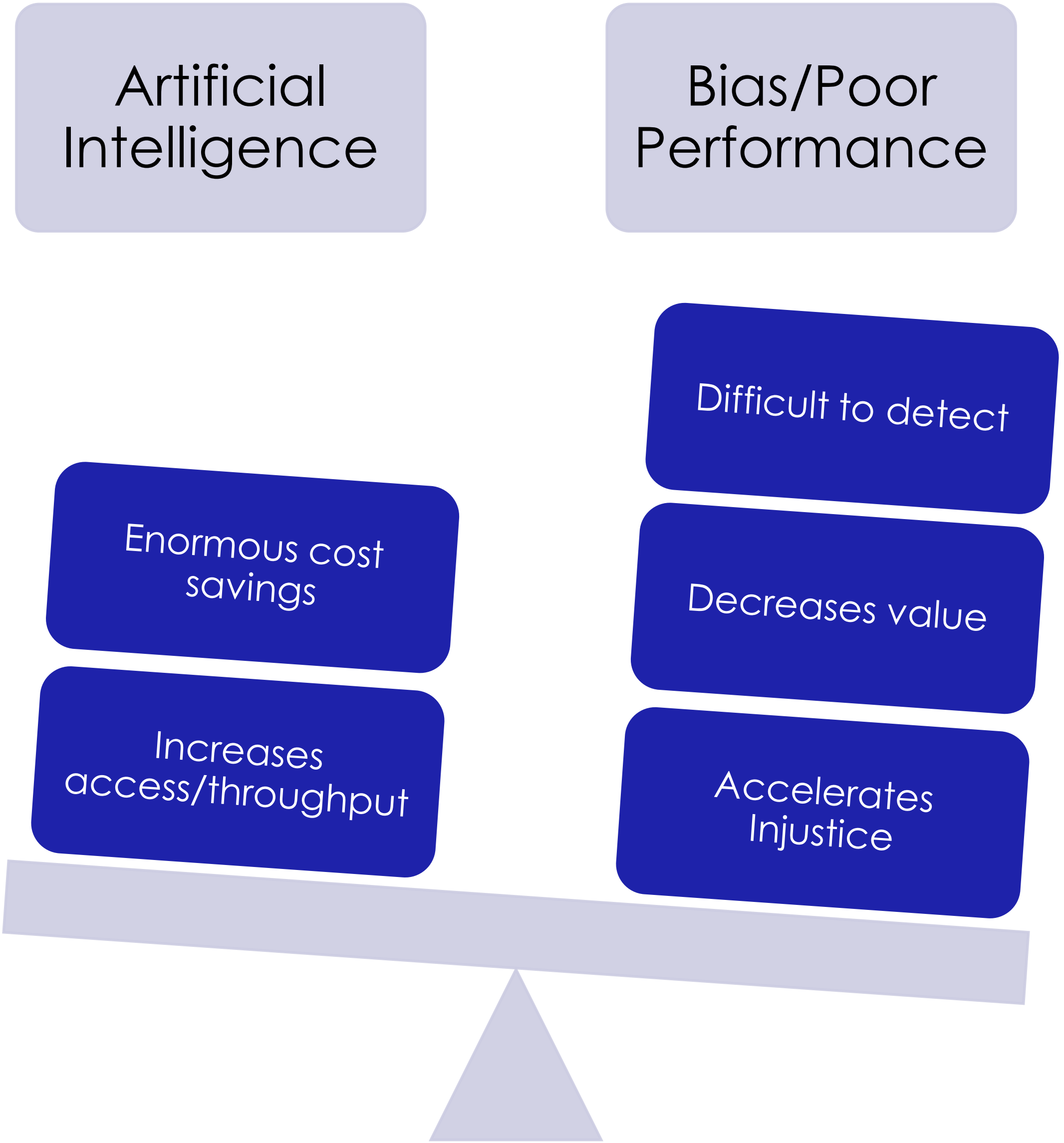
HEALTH

A biased test kept thousands of Black people from getting a kidney transplant. It's finally changing

Jazmin Evans had been waiting for a new kidney for four years when her hospital revealed shocking news: She should have been put on the transplant list in 2015 instead of 2019 _ and a racially biased organ test was to blame. (AP Video by Tassanee Vejpongsa; Production by Shelby Lum)

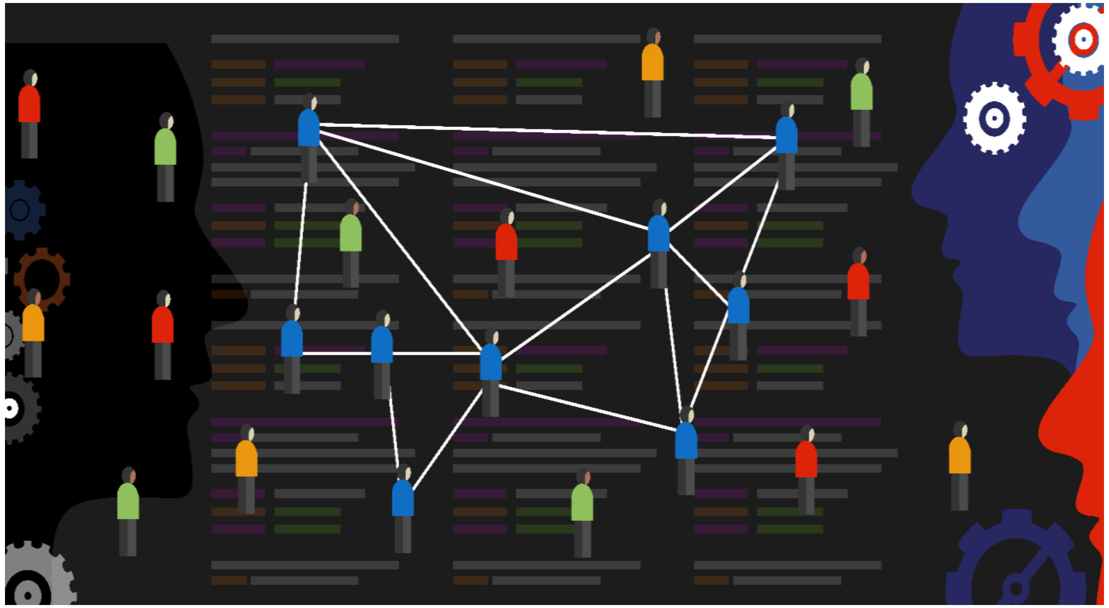
Photos 3

Neergaard, Lauran. A biased test kept thousands of Black people from getting a kidney transplant. It's finally changing. AP News. Published April 1, 2024.


Artificial Intelligence
Bias/Poor Performance
Enormous cost savings
Increases access/throughput
Difficult to detect
Decreases value
Accelerates Injustice

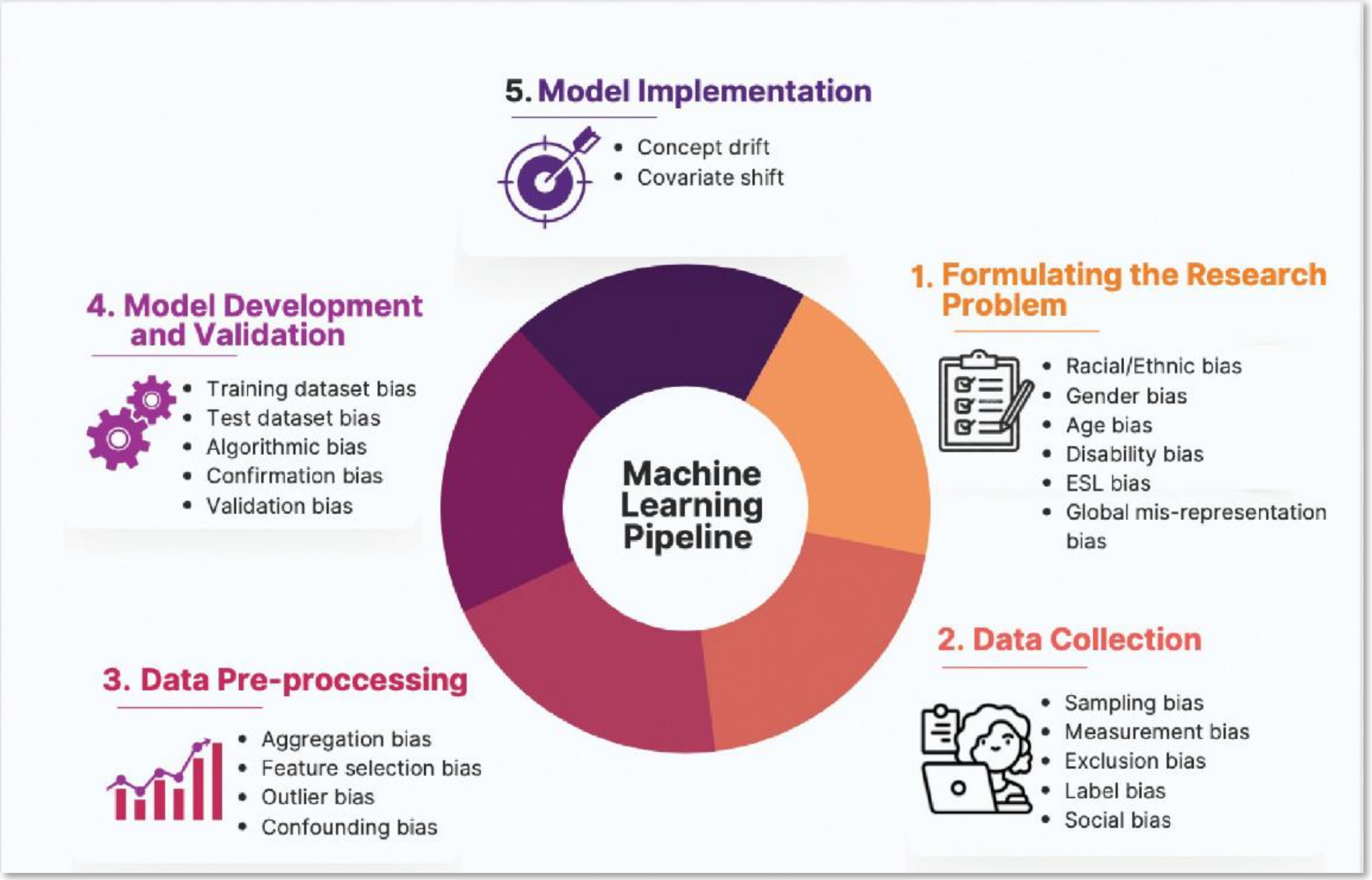
What is Bias in Clinical Algorithms?

Bias refers to the difference in how one or more subgroups is treated, represented or perceived, resulting in unfair/unjust outcomes.



Accessed on July 25, 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Sources of Model Bias

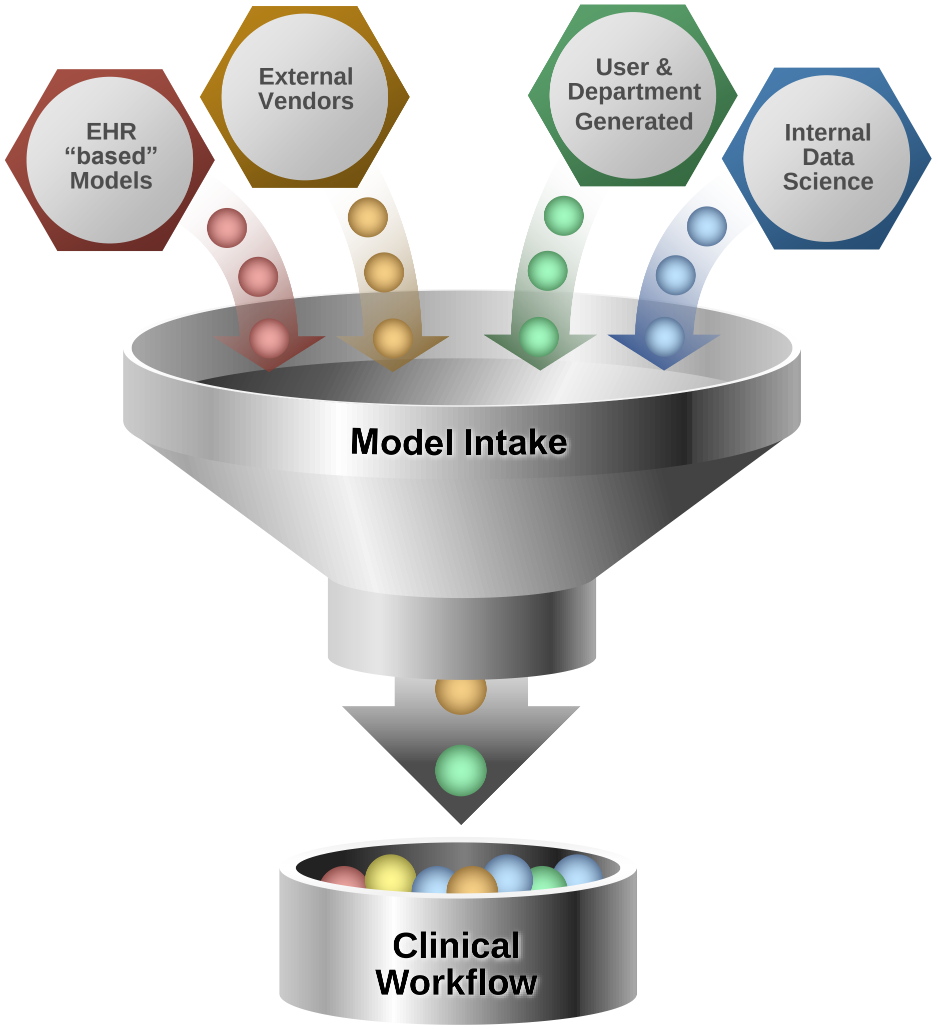


PLOS Digit Health. 2023;2(6):e0000278.

Real World Use

Different:

- Skills
- Knowledge bases
- Resources available
- Make up of project teams



Institutional Governance

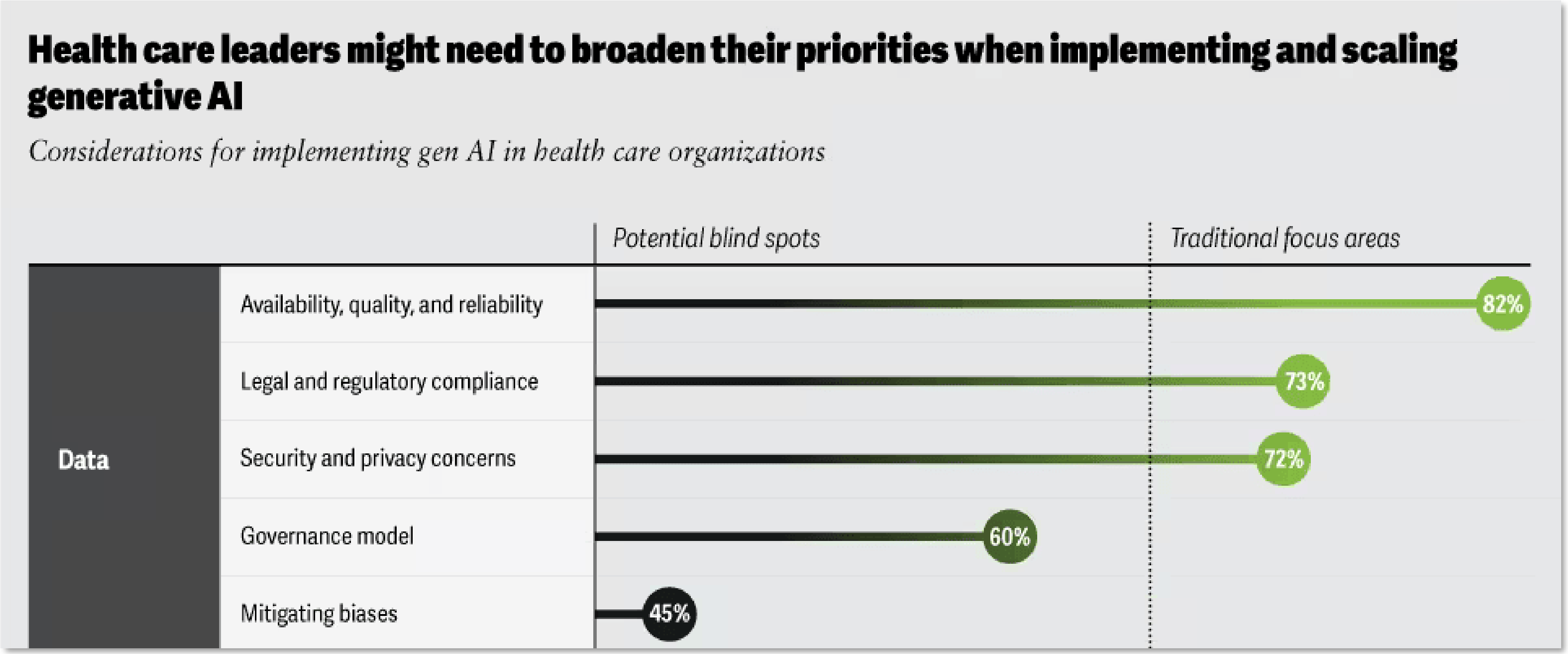
Prediction Models — Development, Evaluation, and Clinical Application

Michael J. Pencina, Ph.D., Benjamin A. Goldstein, Ph.D., and Ralph B. D'Agostino, Ph.D.

*"Given the number of emerging prediction models and their diverse applications, **no single regulatory agency can review them all**. This limitation, however, does not absolve models' developers and users from applying the utmost scrutiny in demonstrating effectiveness and safety."*

N Engl J Med. 2020 Apr 23;382(17):1583-1586.

Deloitte Survey of 60 Healthcare Leaders



Overcoming generative AI implementation blind spots in health care. Deloitte Insights. Accessed March 6, 2024

ABCDS Mission Statement

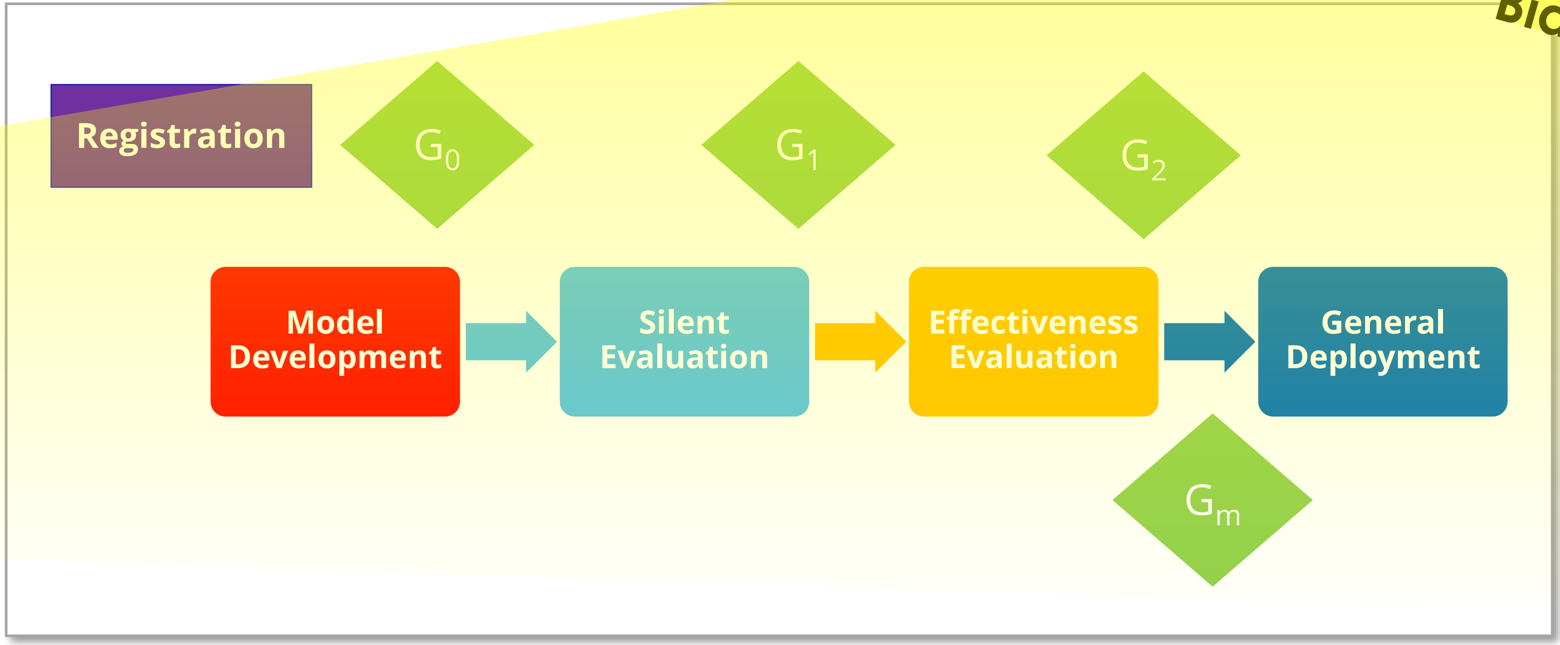
*"Out of our primary focus on patient safety and high-quality care, our mission is to guide **algorithm-based clinical decision support (ABCDS)** tools through their lifecycle by providing governance, evaluation, and monitoring."*

Principles for Responsible AI

- Define the task we want the AI tool to accomplish
- Describe what success and harm look like
- Create transparent systems for continuously testing and monitoring AI tools
- Ensure that AI technology serves humans

Mitigating Algorithmic Bias Through Oversight

ABCDS Oversight process for the governance, evaluation and monitoring of algorithms to be deployed at Duke Health

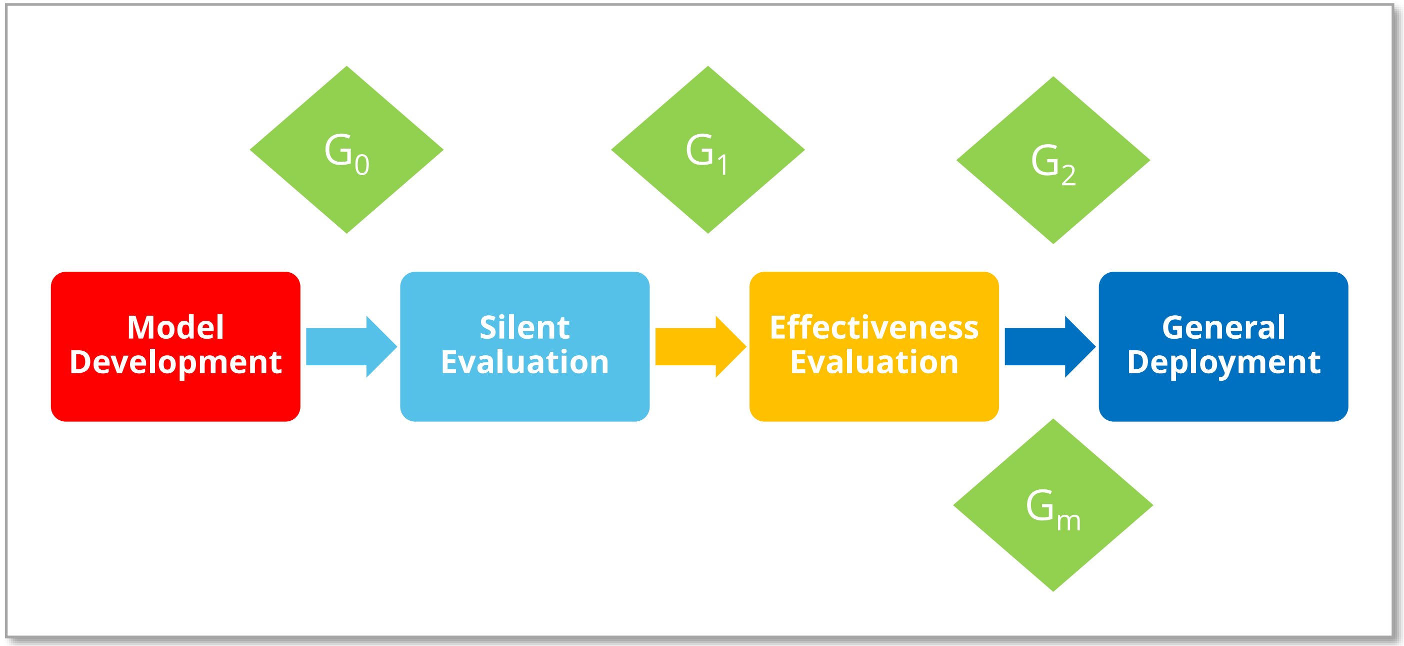


Quality & Ethical Principles

Transparency & Accountability
Impact & Safety
Fairness & Equity
Usability & Adoption
Regulatory Compliance

Process

ABCDS Lifecycle & Our Framework



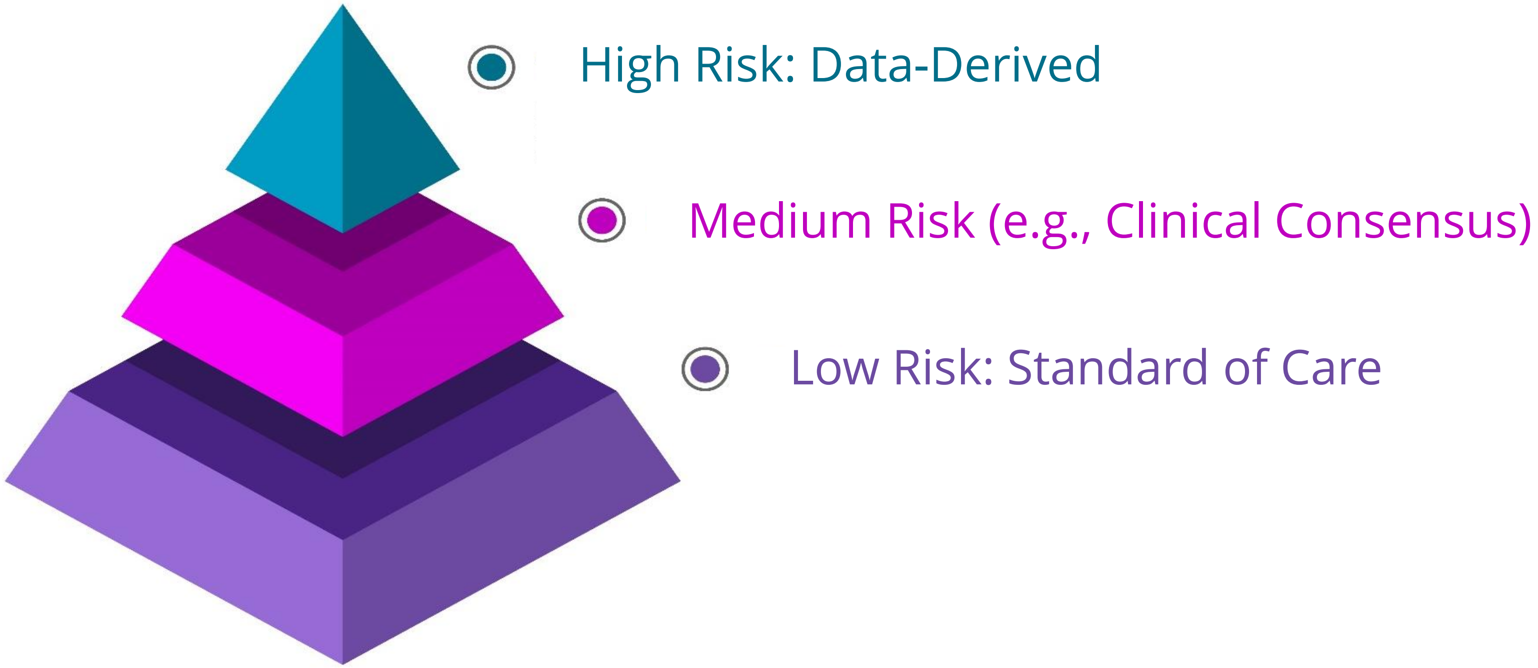
'Just-in-time' Check-Points (**G**ates) Help Model Owners Get Ready for What's Ahead

- What are the clinical outcome and performance metrics?
- How has the model been evaluated?
- Who is the Clinical Owner?
- Who will cover maintenance costs in production?
- Will this ABCDS tool be used outside of Duke Health?
- Is this a standard of care model?
- How will the model be used in the clinic and how is it integrated with the workflow?

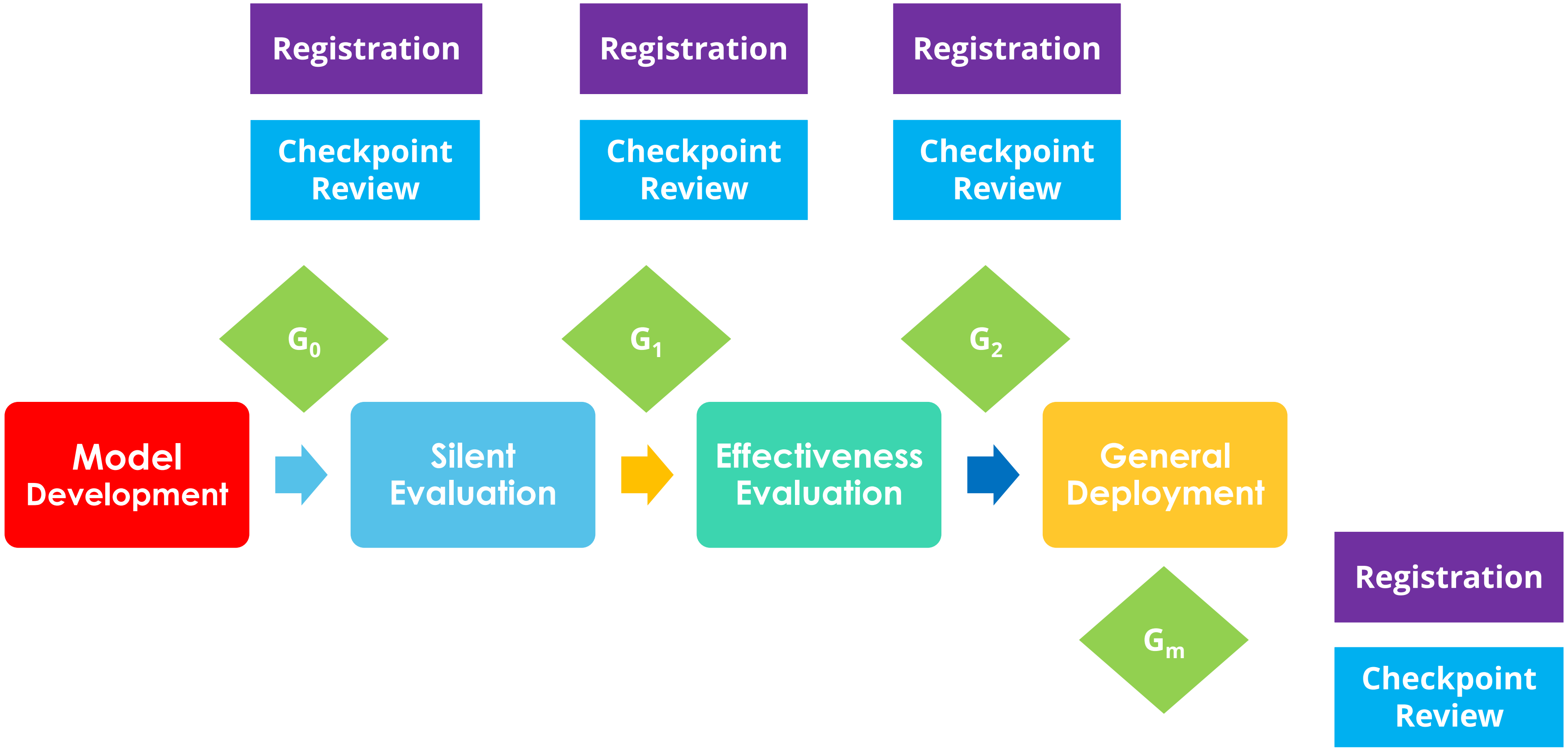
Scope of ABCDS Oversight Framework

ABCDS Tool = Algorithm(s) + Interface Algorithms Are Presented In

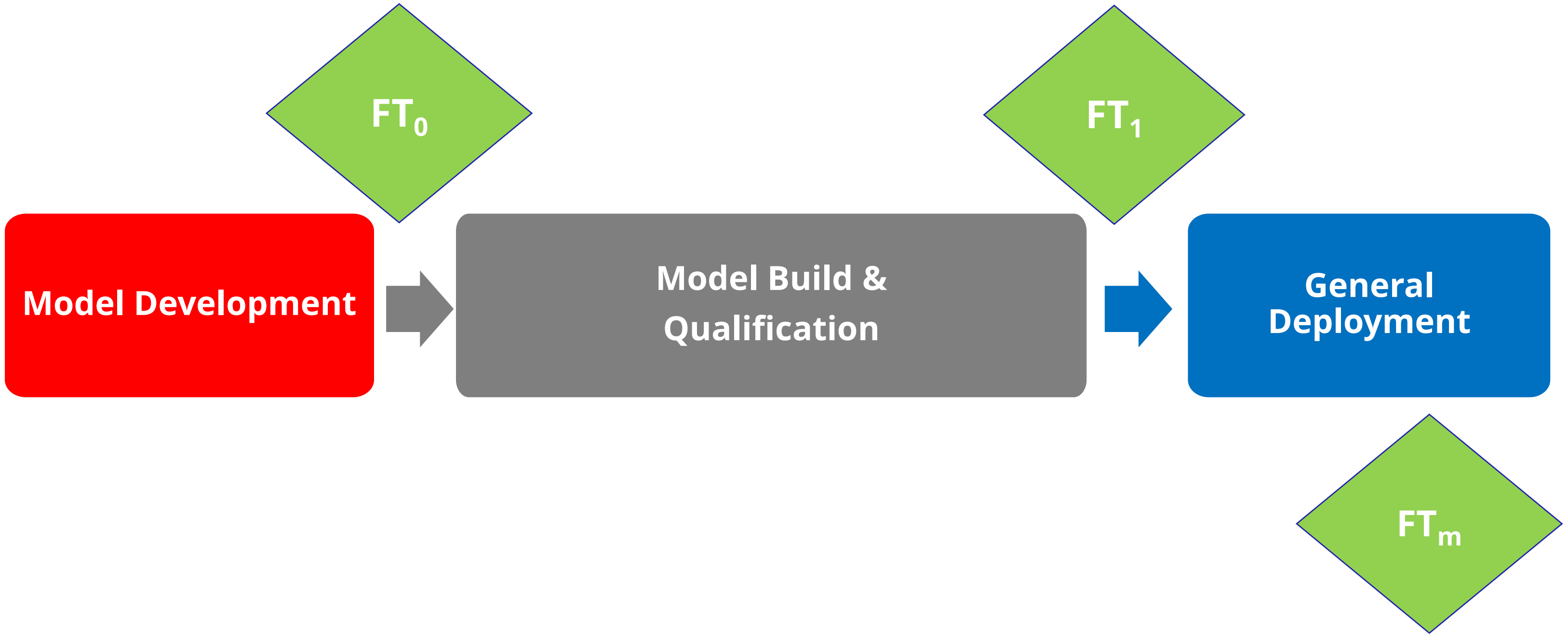
All electronic algorithms that could impact patient care at Duke Health fall within the scope of the ABCDS Oversight Committee and must undergo registration.



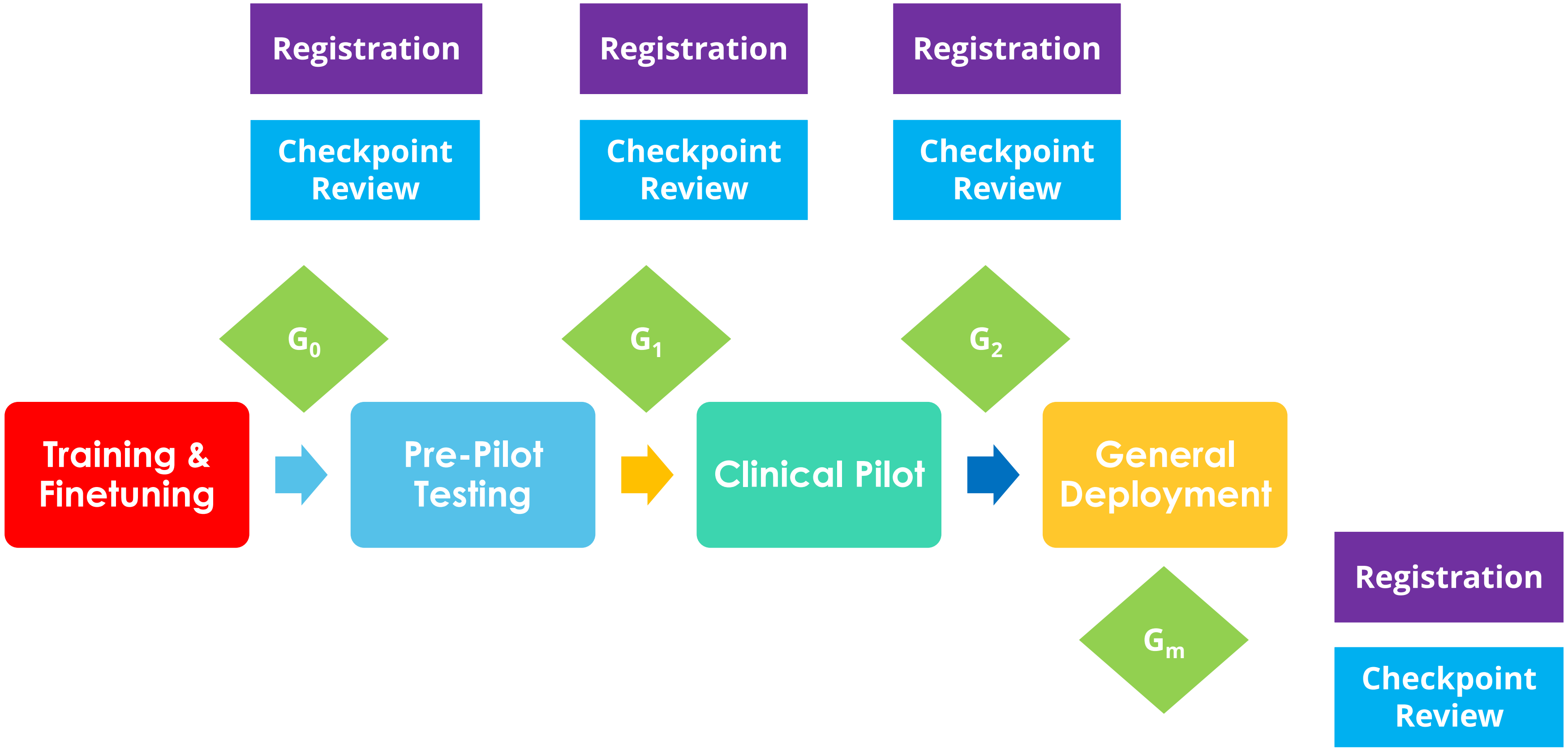
Full Checkpoint Reviews - Predictive



“Fast Track” Checkpoint Reviews



Full Checkpoint Reviews - LLM



Implementing Quality & Ethics with Our Framework

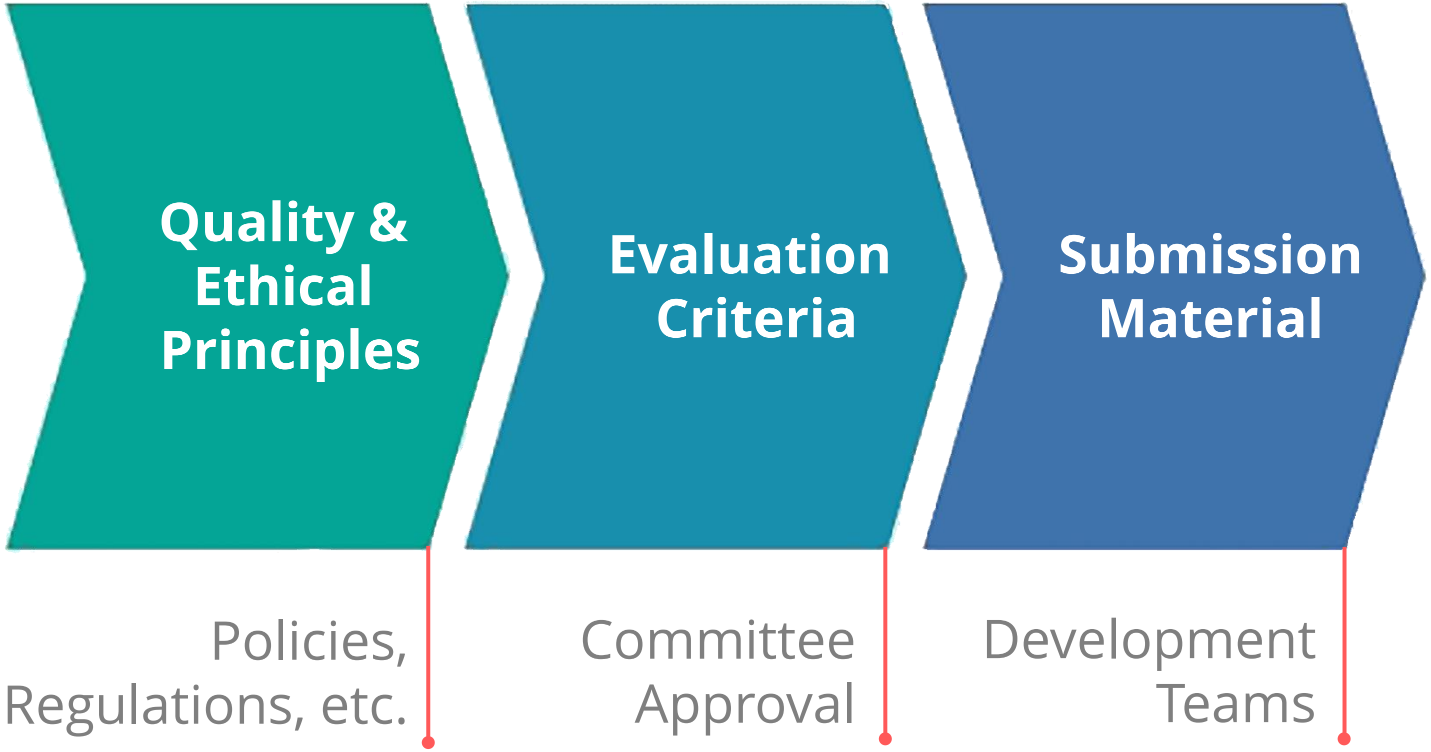
Transparency & Accountability

Impact & Safety

Fairness & Equity

Usability & Adoption

Regulatory Compliance



Implementing Quality & Ethics with Our Framework

Transparency & Accountability

Impact & Safety

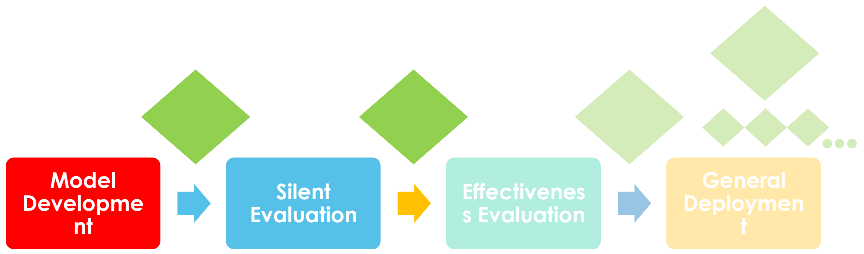
Fairness & Equity

Usability & Adoption

Regulatory Compliance

Principle	Criteria	Submission Materials
Clinical Impact & Safety	The ABCDS software, in comparison to current state, stands to improve clinical care.	✓ Evidence that the tool has potential to impact clinical outcomes or processes ✓ List of key impact metrics (clinical outcomes and/or process improvement) with definitions, following TRIPOD guidelines[5] ✓ List of core performance metrics (e.g. sensitivity, PPV, etc.) and results from development ✓ Calibration curves, threshold selections and justification if applicable
	Plans for Silent Evaluation will inform the decision to proceed with pilot implementation in clinic.	Silent Evaluation Plan, including: ✓ Summary of benefits you expect to demonstrate and criteria to proceed into Effectiveness Evaluation ✓ Study design, including in/exclusion criteria, timeframe and sample size considerations ✓ Core performance metrics with shell tables ✓ Data analysis plan ✓ Data quality evaluation plan

Sample evaluation criteria supporting the principle of clinical impact and safety at the G_0 Checkpoint evaluation between pilot implementation and general deployment



(Unpublished work)

Adapting to Generative AI

McKinsey on Generative AI in Healthcare

*"Gen AI represents a meaningful new tool that can help unlock a piece of the **unrealized $1 trillion of improvement potential** present in the industry."*

Generative AI in healthcare: Emerging use for care | McKinsey. Accessed March 6, 2024.

Ambient Digital Scribes

NEJM Catalyst | Innovations in Care Delivery

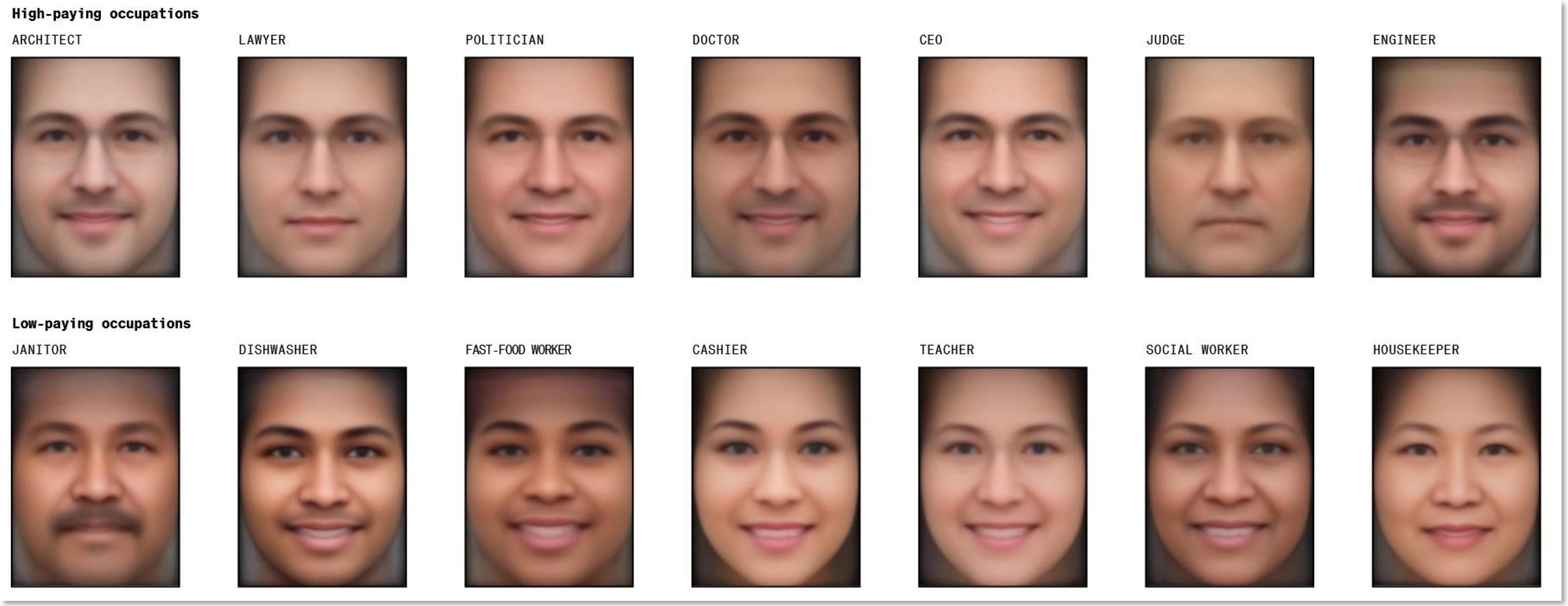
COMMENTARY

Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation

Aaron A. Tierney, PhD, Gregg Gayre, MD, Brian Hoberman, MD, MBA, Britt Mattern, MBA, Manuel Ballesca, MD, Patricia Kipnis, PhD, Vincent Liu, MD, MS, Kristine Lee, MD

Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation | NEJM Catalyst. Accessed August 27, 2024. https://catalyst.nejm.org/doi/full/10.1056/CAT.23.0404

Generative AI Poses New Risks



Nicoletti, Leonardo, Bass, Diana. Generative AI Takes Stereotypes and Bias From Bad to Worse. *Bloomberg.com*. https://www.bloomberg.com/graphics/2023-generative-ai-bias/. Accessed February 22, 2024.

Different from Predictive Models

- Performance Metrics of “traditional” machine learning may not apply
  - Precision/Recall
  - Accuracy
  - F1 score
- Corollaries in NLP (BLEU, ROUGE, etc.) may not align with human evaluation
- May not have a “fully silent” evaluation

Evaluation Options for LLM Output

Human Evaluation

- Captures nuance, direct feedback
- Costly, may have biases and inconsistencies

Intrinsic Metrics (e.g., BLEU, ROUGE, BERTScore)

- Reproducible, large-scale evaluation
- May not correlate with human evaluation

Task-specific Benchmarks (e.g., GLUE, SuperGLUE)

- Consistent, reproducible
- Limited to predefined tasks with well-established metrics

DeLucia S. Using LLMs To Evaluate LLMs. Arize AI. Published January 16, 2024. Accessed March 5, 2024. https://medium.com/arize-ai/using-llms-to-evaluate-llms-c69da454048c

Steps for Human Evaluation

1. Assemble the Team of Stakeholders and Experts
2. Develop Evaluation Metrics
3. Choose and Train Pilot Testers
4. Update Metrics based on Experience/Feedback
5. Monitor over Time

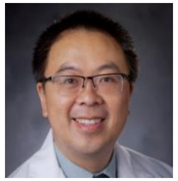
1) Skills/Perspectives Needed

- Qualitative Research, Questionnaire Design, Evaluation
- Data Science
- Medical Ethics
- Clinical Informatics
- End Users
- Patients, depending on the use case

People: ABCDS Oversight Committee

Co-Chairs:


M Pencina


E Poon

Director:


N Economou

ABCDS Oversight Committee

Additional Committee Members:


S Balu


M Cary


M Lipkin


K Lytle

ABCDS Regulatory Advisory Subcommittee

ABCDS Evaluation Subcommittee

ABCDS Implementation and Monitoring Subcommittee

Ops Team::


S Bessias


N Walden

Co-Chairs:

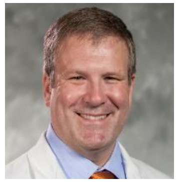

A Parrish


S Elengold

S Ellison

Co-Chairs:


B Goldstein


E Jelovsek

Co-Chairs:


A Bedoya


C O'Brien

2) Develop Evaluation Metrics

- **Qualitative Research: Identify constructs and operationalize them**
- What are the ***virtues*** and ***potential harms*** of your application?
- How will you detect if these are present?
- Translate this into a question for an evaluator
- Remember questionnaire best practices!
  - Framing, Randomization, question burden

Bhandari P. Operationalisation | A Guide with Examples, Pros & Cons. Scribbr. Published May 6, 2022. Accessed March 7, 2024. https://www.scribbr.co.uk/thesis-dissertation/operationalisation/

Validated Frameworks: MQM

Minor Errors are technically errors, but do not disrupt the flow or hinder comprehension.

Major Errors disrupt the flow, but what the text is trying to say is still understandable.

Critical Errors inhibit comprehension of the text.

Accuracy: If there is an error with the translation, that has to do with the fact that it is a translation, try to place it in a category below. If it doesn't match any of those categories, place it here as a general Accuracy error.

Terminology: The word is correct, but not the one usually used in that domain.

Example- Using 'Large shallow pan' as opposed to 'sauté pan.'

Mistranslation: Something has been mistranslated.

Example- *Il* being translated as 'he' instead of 'it.'

Untranslated: Something is still in French.

Note: Proper nouns should stay in French!

Omission: Something is missing from the translation.

Example- A word, phrase or sentence is left out entirely.

Addition: Information has been added.

Example- The translator has added 'a city in France' after 'Paris.'

Fluency: If there is an error related to the text that would still be an error if the text were not a translation, try to place it in a category and sub-category below. If it does not match any of those categories, place it here as a general Fluency error.

Content: The error is related to the content of the text. If it fits into a subcategory please put it there.

Unintelligible: The text makes no sense, but the error does not fall into another category.

Example- 'ao;sdtnq'

Mechanical: A problem with the mechanics/presentation of the text. If the error fits into a subcategory please put it there.

Inconsistency: The text has inconsistent information.

Example- Lists the due date as two different dates, a location as both to the east and west.

Register: The text is too formal or too informal.

Note: This is a newspaper article; so that level of formality.

Style: The style of the text does not feel like a newspaper.

Example- Sentences are correct, but simply too long.

Locale Convention: Uses a word from the wrong locale.

Example- Using a Canadian word in a translation for France.

Spelling: A word is misspelled.

Note: This includes missing accent marks.

Typography: Errors in punctuation and other keyboard errors.

Example- Extra spaces, missing commas, un-capitalised letters.

Grammar: Error in grammar or syntax that is not spelling or typography.

Example- 'him house' vs. 'his house.'

Mariana VR. The Multidimensional Quality Metric (MQM) Framework: A New Framework for Translation Quality Assessment. In: ; 2014. Accessed September 27, 2023. https://www.semanticscholar.org/paper/The-Multidimensional-Quality-Metric-(MQM)-A-New-for-Mariana/9ac4cca8f64bd4c1e5fcc23af9b5b1b84bdc0774

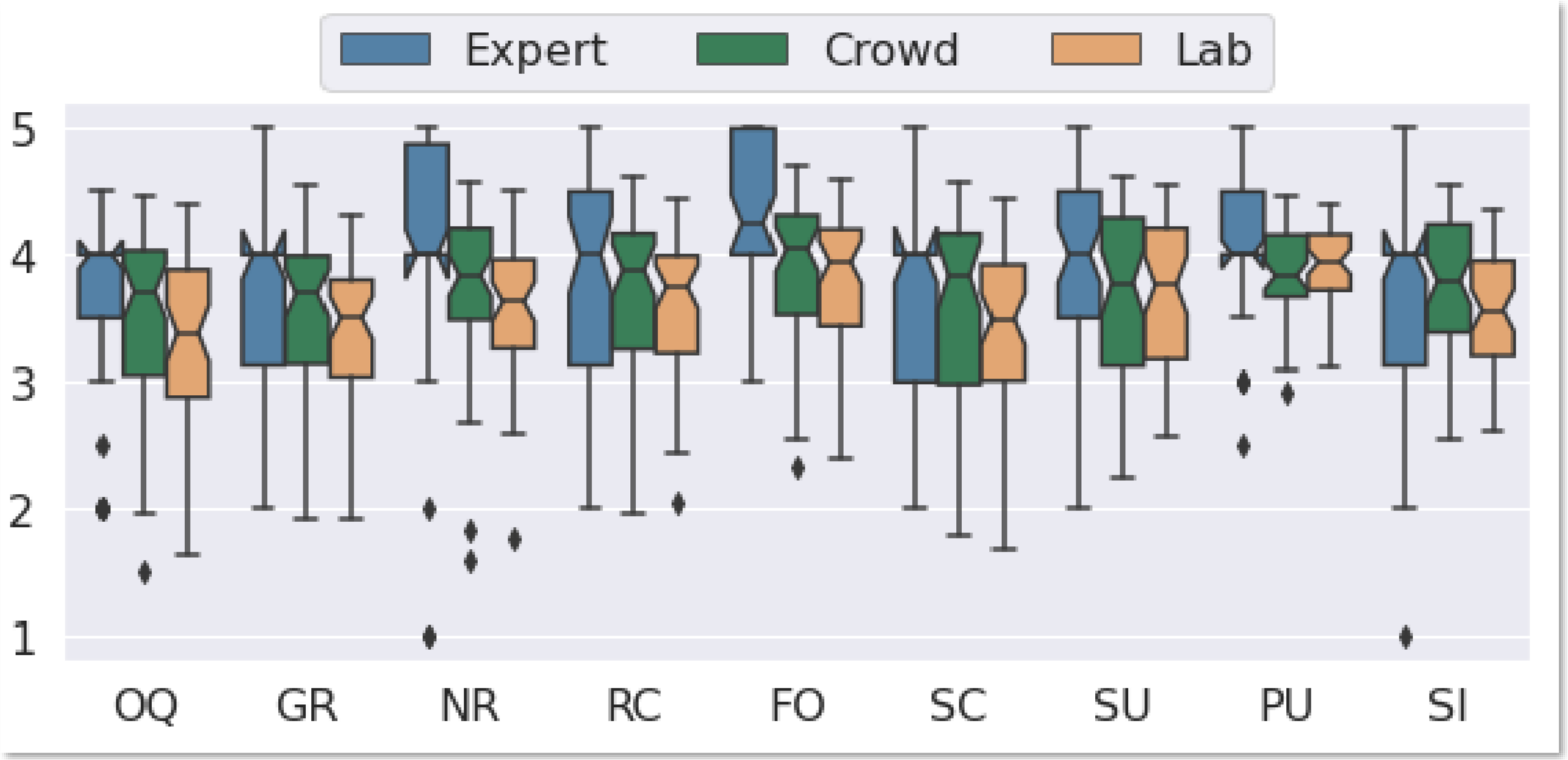
PDQI-9: EHR Notes

Attribute	Score	Description of Ideal Note
1. Up-to-date	Not at all 1 2 3 4 Extremely 5	The note contains the most recent test results and recommendations.
2. Accurate	Not at all 1 2 3 4 Extremely 5	The note is true. It is free of incorrect information.
3. Thorough	Not at all 1 2 3 4 Extremely 5	The note is complete and documents all of the issues of importance to the patient.
4. Useful	Not at all 1 2 3 4 Extremely 5	The note is extremely relevant, providing valuable information and/or analysis.
5. Organized	Not at all 1 2 3 4 Extremely 5	The note is well-formed and structured in a way that helps the reader understand the patient's clinical course.
6. Comprehensible	Not at all 1 2 3 4 Extremely 5	The note is clear, without ambiguity or sections that are difficult to understand.
7. Succinct	Not at all 1 2 3 4 Extremely 5	The note is brief, to the point, and without redundancy.
8. Synthesized	Not at all 1 2 3 4 Extremely 5	The note reflects the author's understanding of the patient's status and ability to develop a plan of care.
9. Internally Consistent	Not at all 1 2 3 4 Extremely 5	No part of the note ignores or contradicts any other part.

Appl Clin Inform. 2012;3(2):164–174.

3) Choose and Train Pilot Testers

- Testers should align with end users in terms of workflow context and expertise
- Plan to train users in how to perform evaluation



Iskender N, Polzehl T, Möller S. Proceedings of the workshop on human evaluation of NLP systems (HumEval). 2021. Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead.

Training Effect on Experts

	Before Mediation				After Mediation			
	Crowd Summ.		TextRank Summ.		Crowd Summ.		TextRank Summ.	
	Agr. in %	κ	Agr. in %	κ	Agr. in %	κ	Agr. in %	κ
OQ	54	0.228	22.2	-0.040	82	0.637	85.2	0.717
GR	42	0.078	18.5	0.086	78	0.626	88.9	0.809
NR	34	-0.012	11.1	-0.084	70	0.520	85.2	0.797
RC	56	0.381	29.6	0.013	88	0.819	92.6	0.882
FO	52	0.249	88.9	0.779	80	0.685	96.3	0.922
SC	42	0.212	22.2	0.070	82	0.743	85.2	0.783
SU	44	0.220	37	0.093	76	0.635	88.9	0.839
PU	38	0.005	48.1	0.169	70	0.469	92.6	0.856
SI	34	-0.038	40.7	0.234	78	0.565	92.6	0.886

Iskender N, Polzehl T, Möller S. Proceedings of the workshop on human evaluation of NLP systems (HumEval). 2021.
Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead.

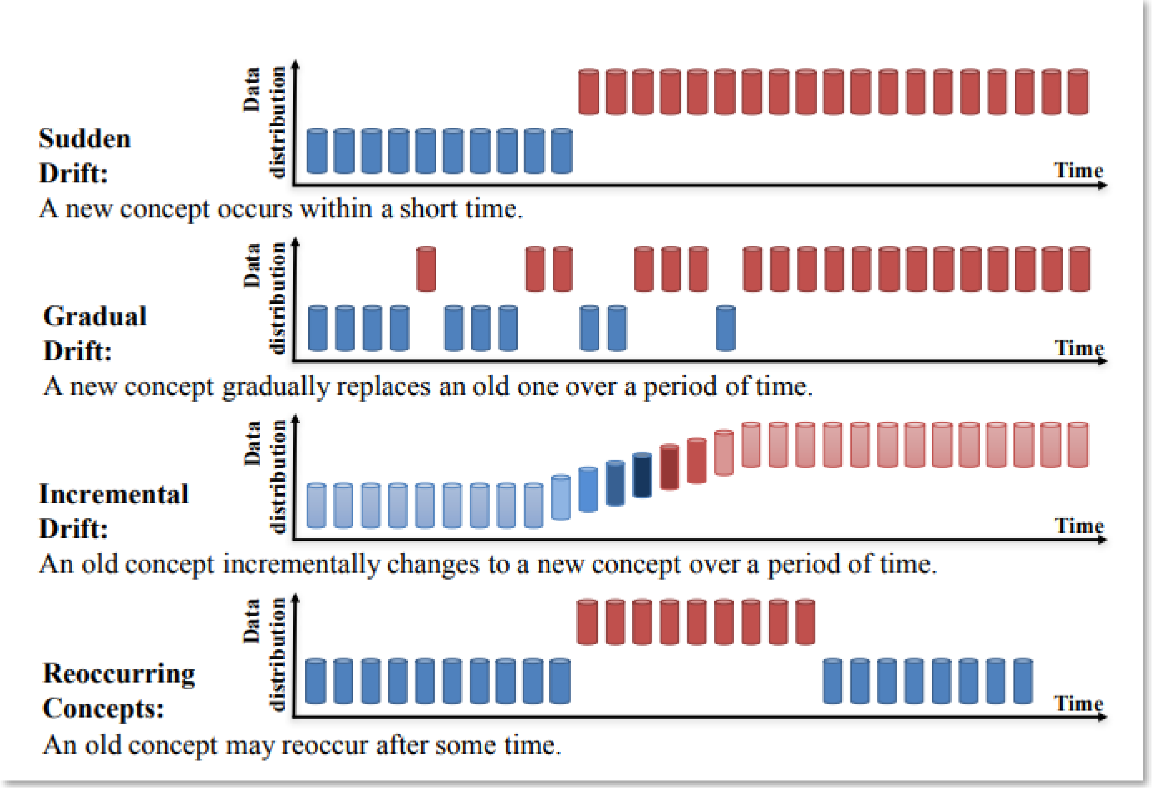
4) Monitor for Drift over Time

"G_M" in ABCDS process

Plan for periodic re-evaluations

Changes in

- Foundation Model
- Prompt
- Data pre-processing
- Model input distribution



Use Cases at Duke

LLM-generated In Basket Drafts

Silent Evaluation

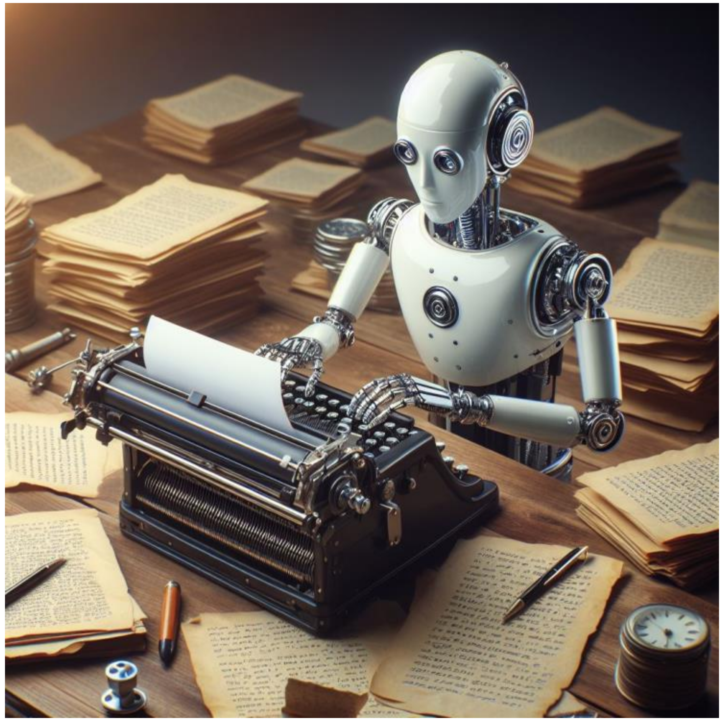
- 200 de-identified and synthetic messages for prompt engineering
- 100 out-of-sample messages
- Five informaticians evaluated
- “Would you use this reply with minor modifications?”
    - If not, categorize the reason for failure
- Expansion to pilot users for prompt engineering

G_0 Process and Evaluation

Would you use this reply with minor modifications instead of composing the entire reply yourself?

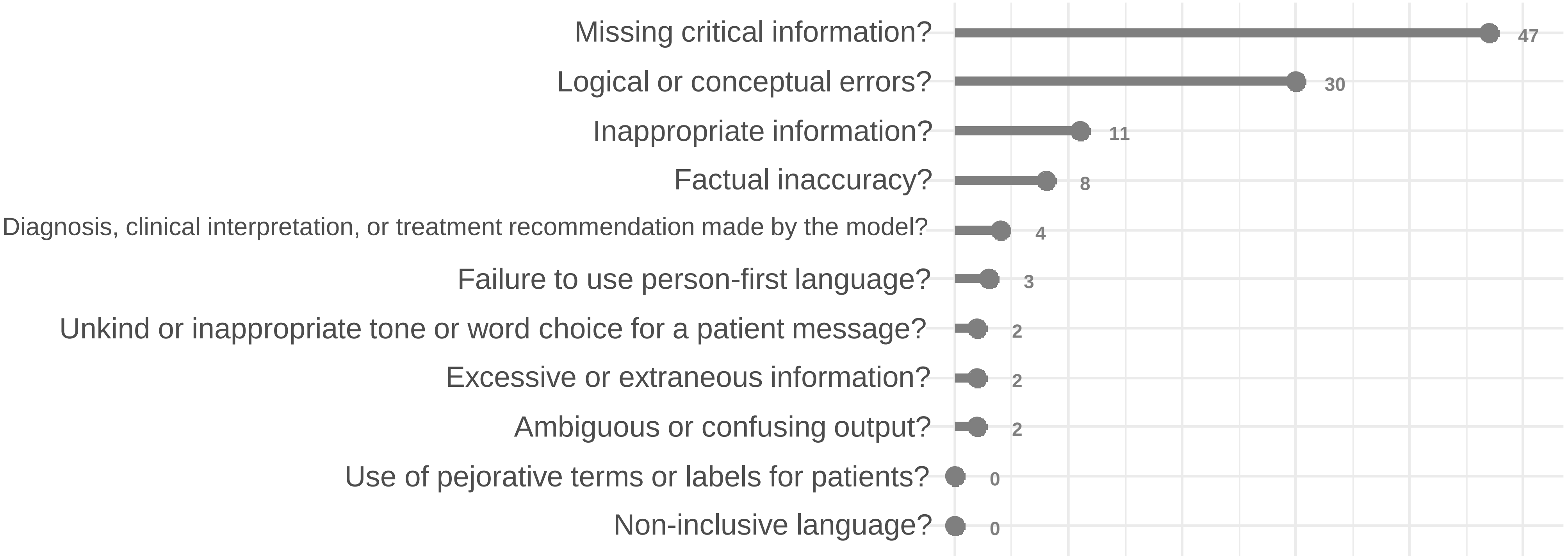
412 total evaluations (overlap)

Category	Pass Rate
General	72.65%
Paperwork	91.84%
Refills	75.42%
Results	84.38%
Overall	79.37%



Generated with AI · March 20, 2024 at 11:43 AM
Microsoft Designer, DALL-E 3

Failure Reasons



Ambient Digital Scribes

Phase 1: Safety Review

- 20 evaluators across 16 specialties
- 2-4 weeks
- Safety Evaluation of ~200 notes

Phase 2: Value Proposition, Expansion Planning

- Aiming for 300-500 users
- 60-90 days
- Pre-post surveys
- Biweekly feedback sessions
- **Informs future user onboarding**

Safety Evaluation (Likert; modified PDQI-9)

- **Take notes during the visit to compare to the draft**
  - Critical omissions?
  - Hallucinations/factual inaccuracies?
  - ***Stigmatizing language?***
  - Misleading information?
  - Grammar/style/tone errors?
- **How difficult were these to find and correct**
- **Estimate the degree of harm this could cause if left uncorrected.**

User Training: Stigmatizing Language

Home > Journal of General Internal Medicine > Article

How to Reduce Stigma and Bias in Clinical Communication: a Narrative Review

Narrative Review | Published: 06 May 2022

Volume 37, pages 2533–2540, (2022) Cite this article

Download PDF

Journal of General Internal Medicine

Aims and scope →

Submit manuscript →

Megan Healy MD, Alison Richard BA & Khameer Kidia MD

J Gen Intern Med. 2022;37(10):2533–2540.

Language to Detect

- Disease-First Language
- Pejorative Terms
- Non-Inclusive Language
- Labels
- Weaponized Quotations
- Race or Socioeconomic framing
- Blame/Judgement
- Undermining the patient's experience

Review: Safety and Bias

2 categorical questions:

- Visit Medium (in person or virtual)
- Visit Type

10 multiple choice questions assessing:

- 1) Quality of CoPilot-produced note (error categories from PDQI9)
- 2) Harm that might occur if note was not altered by provider

2 open feedback questions

- Additional observations
- Concerns about the technology

Appl Clin Inform. 2012;3(2):164–174.

Survey Structure: Scoring Evaluation

Effort required to correct note:

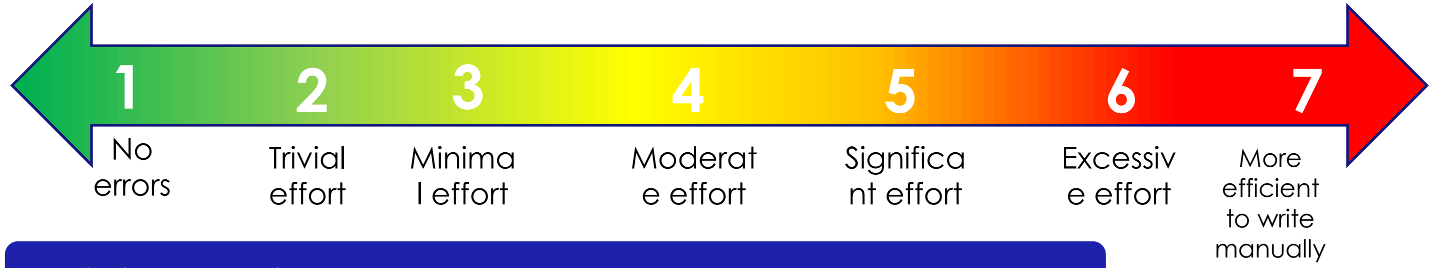
- 1 = No errors
- 2 = Trivial effort
- 3 = Minimal effort
- 4 = Moderate effort
- 5 = Significant effort
- 6 = Excessive effort
- 7 = More efficient to manually write note

Harm if errors were uncorrected

- None/NA
- Mild
- Moderate
- Severe

Survey Results (n=216)

Effort required to correct



Missing details

- Effort required to correct: **2.31**
- Harm if uncorrected: 63% none, 29% mild, **8% moderate**

Incorrect statements

- Effort required to correct: **2.10**
- Harm if uncorrected: 63% none, 31% mild, **6% moderate**

Harmful, non-inclusive, or stigmatizing language

- Effort required to correct: **1.09**
- Harm if uncorrected: 97% none, 3% mild

Irrelevant, misleading information

- Effort required to correct: **1.72**
- Harm if uncorrected: 79% none, 19% mild, **2% moderate**

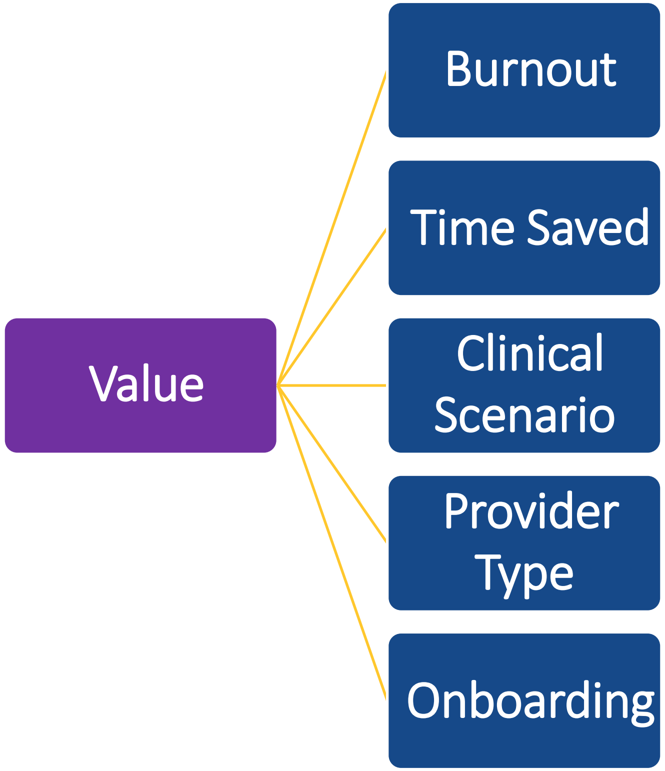
Description of Risk: "Moderate" Concerns

For instances where a **moderate** level of harm could potentially occur, the following concerns were shared:

- Speaker identification including misgendering patient’s spouse
- Incorrect medications dose, changes, discontinuations, name
- Included a medication discussed but decided against
- “The key point of my diagnosis, a serious one, was recording opposite to what I said.”

Value Clarification

- What is the right fit between the tool and situation?
- Surveys
  - Burnout (Copenhagen Burnout Inventory)
  - Cognitive Load (NASA-TLX)
  - Satisfaction/Net Promoter Score
  - Which scenarios created good/bad replies
- Metrics
  - Message composition time
  - Replies outside of work hours
  - Time to chart closure



Lessons Learned

- **Successful AI Governance is a Team Sport**
  - Many skillsets, perspectives and languages to bring together
- **Culture Shift is Hard**
  - Show Teams how to succeed by addressing gaps in their knowledge, skillsets, and/or bandwidth
  - Governance’s role is Coach and Facilitator (not Punisher)
  - There is no such thing as over-communication in a complex system
- **Benefits of Centralized Governance**
  - Transparency of process expectations
  - Institutional visibility into all the ‘skeletons in the closet’
- **Conscious Decision (thus far) Not to Regulate Who Gets to Build AI Models**

Key Lessons for Generative AI

- Success relies on collaboration between governance and operational teams
- Human Evaluation remains the standard for now
- Changes are iterative
- Guidance rather than prescription

Generated with AI · April 11, 2024 at 2:54 PM

Learn More...

https://aihealth.duke.edu/algorithm-based-clinical-decision-support-abcds/

J Am Med Inform Assoc. 2022;29(9):1631–1636.

Contact us at abcds@duke.edu

Special Thanks, Takeaways, and Questions

Special Thanks

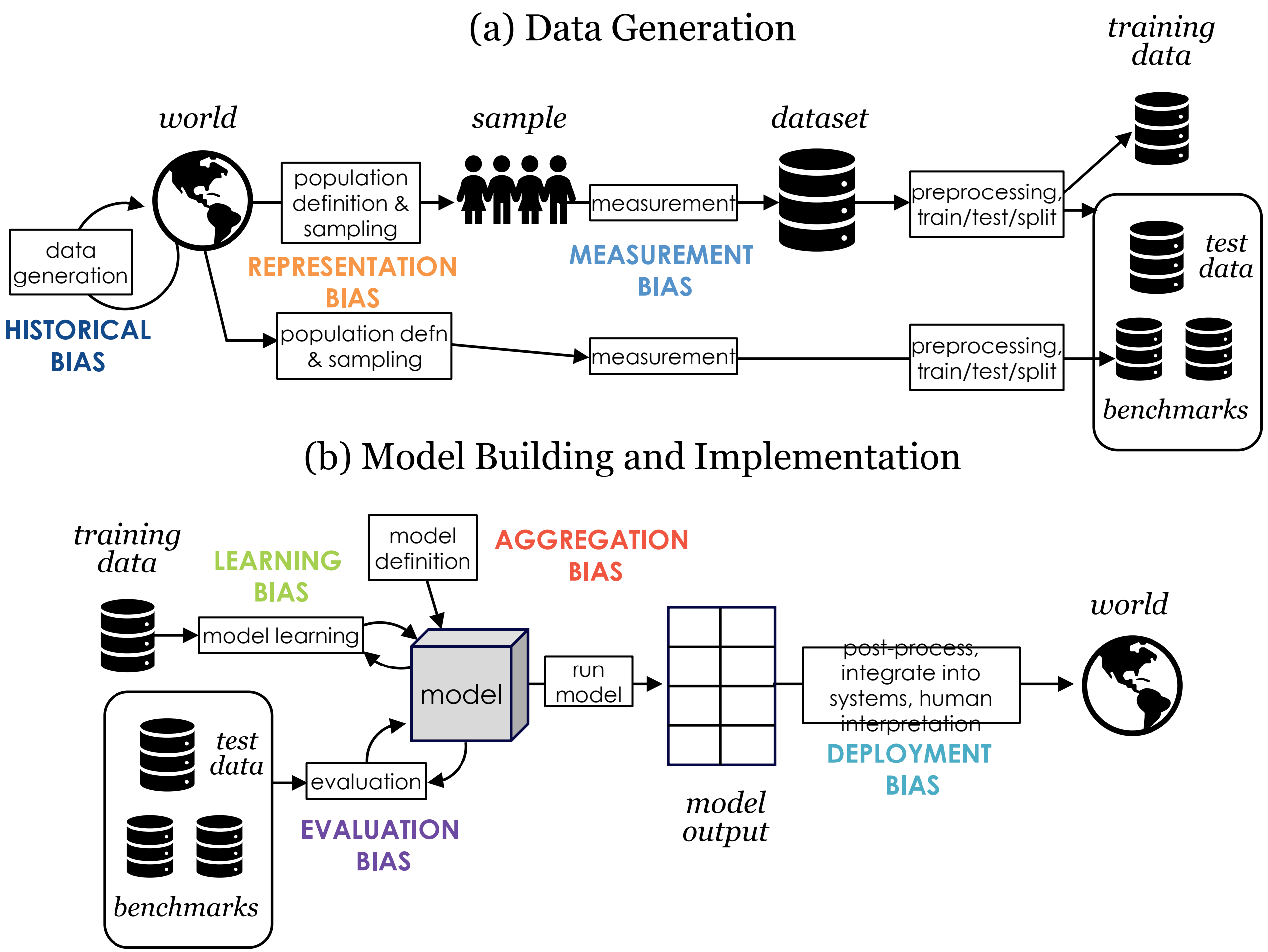
- Eric Poon
- Armando Bedoya
- Anthony Sorrentino
- Jo Cavalier
- Michele Casey
- Chuan Hong
- Sophia Bessias
- Amy Loeblein
- Michael Cary
- Kay Lytle
- Matthew Engelhard
- Nicoleta Economou-Zavlanos
- Karen Ament
- Holland Sink
- Tres Brown III
- Jessica Sperling

Key Takeaways

- Must have collaboration between governance and operational teams
- Human Evaluation remains the standard for now
- Guidance instead of prescription

Contact me: anand.chowdhury@duke.edu

Understanding Sources of Bias



- Societal Bias
- Label Bias
- Aggregation Bias
- Learning Bias
- Representation Bias
- Evaluation Bias
- Human Use Bias

Suresh H, Guttag J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9). doi:10.1145/3465416.3483305.

FDA Guidance 2022

Heavy Focus on:

- Independent review
- Healthcare status and time criticality
- Automation bias
- Workflow
- Display risk vs. options for care

Contains Nonbinding Recommendations

Clinical Decision Support Software

Guidance for Industry and Food and Drug Administration Staff

Document issued on September 28, 2022.

The draft of this document was issued on September 27, 2019.

For a software function to be Non-Device CDS and thus exempt, it must meet all the following four criteria to be excluded from the device definition under section 520(o) of the FD&C Act.

1	Not intended to acquire, process, or analyze a medical image or a signal from an in vitro diagnostic device or a pattern or signal from a signal acquisition system
2	Intended for the purpose of displaying, analyzing, or printing medical information about a patient or other medical information
3	Intended for the purpose of supporting or providing recommendations to an HCP about prevention, diagnosis, or treatment of a disease or condition
4	Intended for the purpose of enabling an HCP to independently review the basis for the recommendations that such software presents so that it is not the intent that the HCP rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient

Algorithm Type Definitions

- ***A data-driven model (non-standard of care)*** is a model that builds relationships between input and output data using statistical/machine learning techniques. ML/AI and other statistically-derived models fall under this category.
- ***A clinical consensus-based (knowledge-based) model*** *is a formula or set of rules that were derived based on clinical acumen and consensus, the literature, and/or expert recommendations. These algorithms provide the same results on the same inputs.*
- ***Medical standard of care*** is typically defined as the level and type of care that a reasonably competent and skilled health care professional, with a similar background and in the same medical community, would have provided under the circumstances. A ***'standard of care' tool or model*** would be a tool or model used to guide standard-of-care as defined above and would be supported by evidence in the medical literature, recommended by medical societies, or incorporated into clinical practice guidelines.

Societal Bias

Bias Type	Example	Assessment	Mitigation Strategy
Societal Bias Bias due to training data shaped by present and historical inequities and their fundamental causes	Predictive policing algorithms[1] are trained on data that reflects structural racism and criminalization of, e.g., homelessness and poverty. Groups that are more likely to interact with the police are more likely to be identified by policing algorithms as “at risk” for future offense.	*Please discuss the real-world inequities reflected in your training data and how they inform the problem formulation and intended purpose of your model.*	• *Restriction to particular settings or use cases* • *Human-in-the-loop deployment design* • *Multi-stakeholder engagement*

- Label Bias
- Aggregation Bias
- Learning Bias
- Representation Bias
- Evaluation Bias
- Human Use Bias

[1] Angwin, J. Larson, S. Mattu, L. Kirchner, “Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks,” *ProPublica*, 23 May 2016; www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Label Bias

Bias Type	Example	Assessment	Mitigation Strategy
Label Bias Use of a biased proxy target variable in place of the ideal prediction target.	An algorithm[1] used to identify patients for high-risk care management services predict healthcare costs as a proxy for healthcare *need*. Despite having greater health needs, Black patients have lower average healthcare spending (due to structural barriers in access to care) and are thus less likely to be recognized by the algorithm as 'high risk.'	*Please discuss any proxies used as inputs or outputs. Provide a rationale and describe implications for use.*	• *Eliminating proxies (where possible) or choosing a proxy as close as possible to the intended idea or concept*

Label Bias
Aggregation Bias
Learning Bias
Representation Bias
Evaluation Bias
Human Use Bias

[1]Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342.

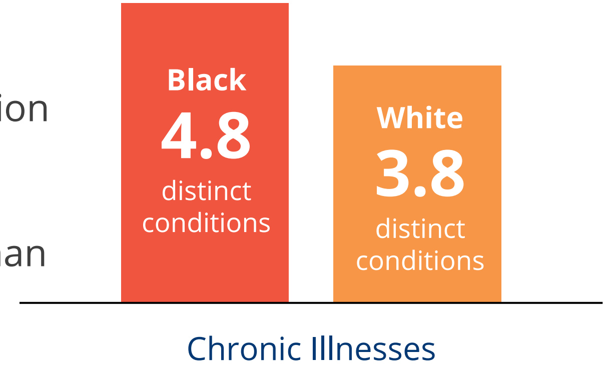
Why is it Important to Identify Racial/Ethnic Bias in Health Algorithms?

Algorithms are used to identify patients with complex health needs in order to provide more comprehensive care management. However, these algorithms can exhibit significant racial bias.

A 2019 study of one such algorithm found:

Black patients who are considerably sicker than White patients are given the same risk score

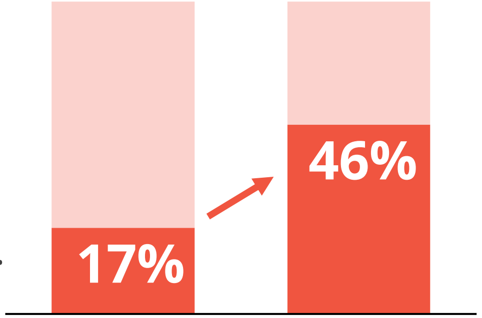
At the risk level that would result in automatic identification for the care management program, Black patients had **26%** more chronic illnesses than White patients.



Why is this?

This algorithm assigned risk scores based on past health care spending. Black patients have lower spending than White patients for a given level of health.

If this bias was eliminated, the percentage of Black patients automatically enrolled in the program would rise from **17%** to **46%**.



[1]Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342.

Aggregation Bias

Bias Type	Example	Assessment	Mitigation Strategy
Aggregation Bias Bias due to use of a one-size-fits-all model for data in which there are underlying groups or types of examples.	A natural language processing (NLP) model developed to scan clinical notes and suggest medication review is used across hospitals in a large health system in which documentation practices differ between locations, leading to poor performance in recently-acquired rural hospitals switching EHR systems.	*Please discuss the ways that the data used to train your model may be observed differently across subgroups.*	• *Use of subpopulation-specific models instead of or in addition to one-size-fits-all models* • *Use of subgroup-specific thresholds in a one-size-fits-all model* • *Imputation or other strategies to improve mapping from inputs to labels across subgroups*

- Label Bias
- Aggregation Bias
- Learning Bias
- Representation Bias
- Evaluation Bias
- Human Use Bias

Learning Bias

Bias Type	Example	Assessment	Mitigation Strategy
Learning Bias Bias due to modeling choices that amplify performance disparities across subgroups.	A development team is working on prediction of asthma exacerbation and uses a variety of methods to generate candidate models. The final model is selected by ranking the candidates on a single performance metric, AUROC. The focus on a single summary metric conceals large performance differences by race leading to poor prediction in the demographic most exposed to environmental asthma triggers.	*Please describe how the model was optimized and the performance metrics used among candidate models.*	• *Penalized optimization methods* • *Subgroup analysis to inform model selection*

Label Bias
Aggregation Bias
Learning Bias
Representation Bias
Evaluation Bias
Human Use Bias

Representation Bias

Bias Type	Example	Assessment	Mitigation Strategy
Representation Bias Bias emerging from non-representative training data which can lead to poor performance in subsets of the deployment population.	A melanoma detection model[1] achieved accuracy parity with a board-certified dermatologist; however, the model was trained primarily on light-colored skin. As such, the algorithm is likely to underperform for patients with dark skin.	*Please discuss the quality and representativeness of your training data.* *If your model is adaptive, please discuss how you will ensure representativeness of the training data on an ongoing basis.*	• *Integration with data from other sources* • *Supplementation with synthetic data* • *Up- or down-sampling approaches* • *Acknowledgement of limitations in model brief or other training materials* • *Refitting an out-of-the-box model to the local population*

- Label Bias
- Aggregation Bias
- Learning Bias
- Representation Bias
- Evaluation Bias
- Human Use Bias

[1]Wang HE, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. *J Am Med Inform Assoc*. 2022 Jul 12;29(8):1323-1333. doi: 10.1093/jamia/ocac065.

Evaluation Bias

Bias Type	Example	Assessment	Mitigation Strategy
Evaluation Bias Bias emerging from a validation dataset that is not reflective of the deployment population and/or the training population.	A health system implements a new vendor model to predict in-hospital deterioration after receiving a validation report showing strong performance in other health systems that share the same EHR. Once the model is connected to the local data source, it produces an unexpected number of false alerts.	*Briefly summarize plans for local validation.*	• *Local validation (required)* • *Re-fitting the model on development sample that better represents the deployment population* • *Post-deployment monitoring with chart review (required)*

Label Bias
Aggregation Bias
Learning Bias
Representation Bias
Evaluation Bias
Human Use Bias

Human Use Bias

Bias Type	Example	Assessment	Mitigation Strategy
Human Use Bias Inconsistent user response to algorithm outputs for different subgroups.	A machine learning algorithm1 developed to help pathologists differentiate liver cancer types did not improve every pathologist's accuracy despite the model's high rate of correct classification. Instead, pathologists' accuracy was improved when the model's prediction was correct but decreased when the model's prediction was incorrect.	*Briefly describe how your algorithm fits into the clinical workflow. If it will replace an existing model or process, please include a comparison to baseline.*	• *Workflow design solutions* • *End user training* • *Post-deployment monitoring with chart review (required)* • *Collection of end user feedback and metrics of adoption*

- Label Bias
- Aggregation Bias
- Learning Bias
- Representation Bias
- Evaluation Bias
- Human Use Bias

Wang HE, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. *J Am Med Inform Assoc*. 2022 Jul 12;29(8):1323-1333. doi: 10.1093/jamia/ocac065.