Bias Testing Considerations for AI Tools in Community Health Centers

April 20, 2026

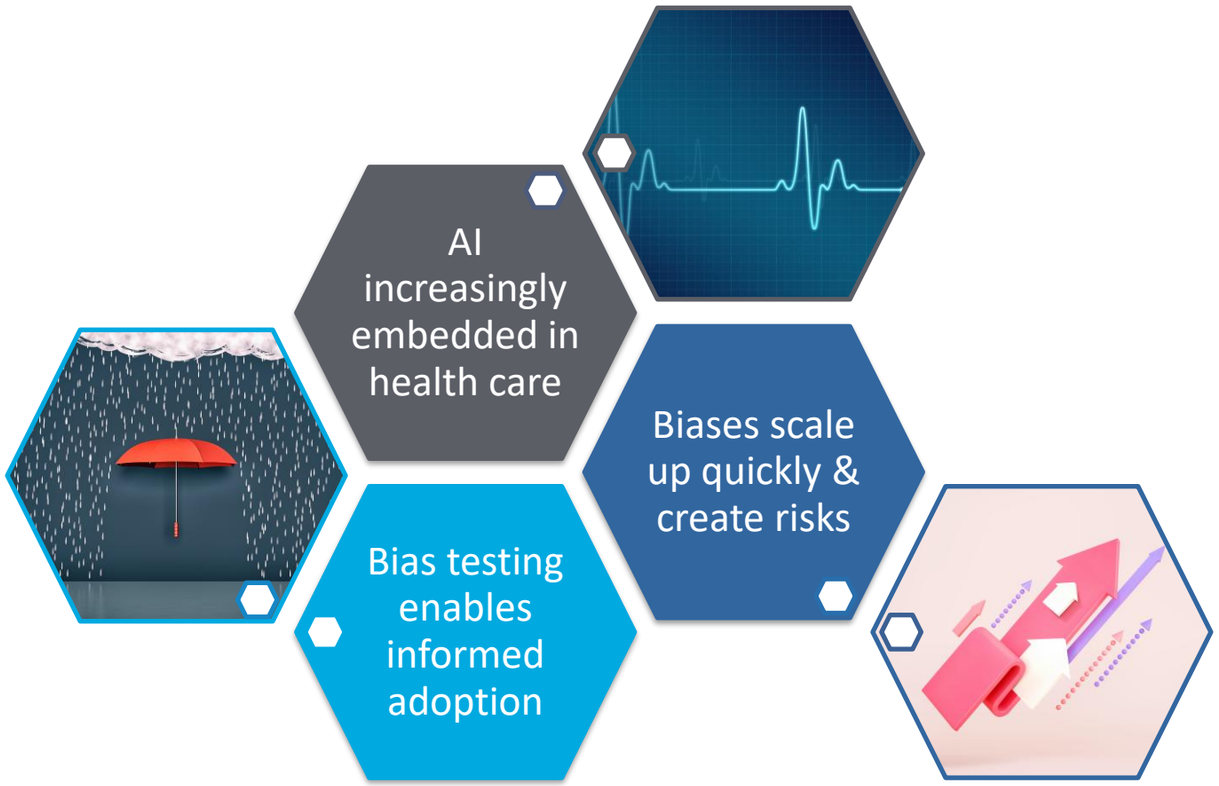
Speakers: Sam Tyner-Monroe, Ph.D.

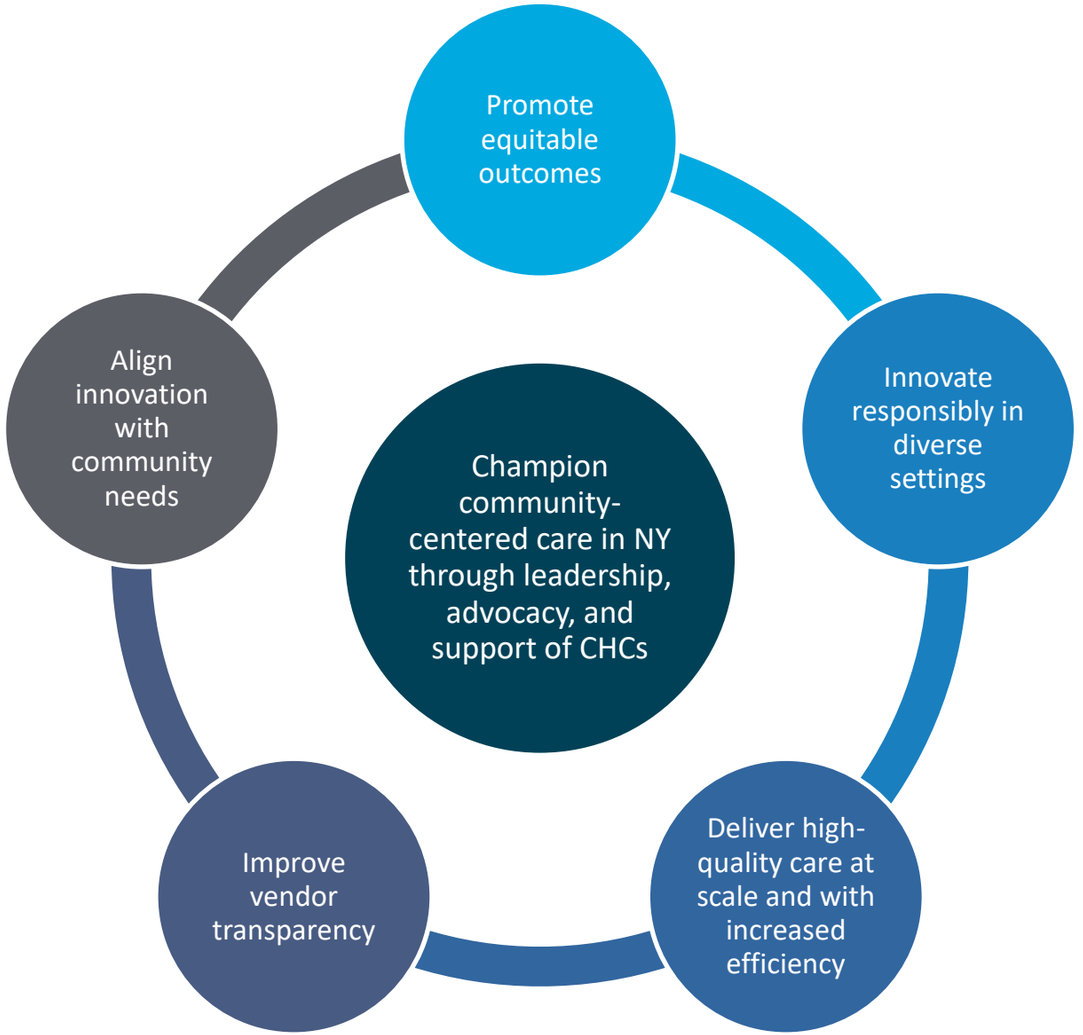
Learning Objectives

- **Recognize** where bias risks arise in commonly procured AI tools used by Community Health Centers (CHCs)
- **Differentiate** high- and low-bias-risk AI use cases in CHC settings and understand what level of testing, oversight, and documentation is reasonable for each.
- **Ask** vendors the right questions upfront, including what evidence, transparency, and safeguards should be expected before adoption.
- **Understand** practical options for monitoring, remediation, and governance that are realistic for resource-constrained organizations.

manatt

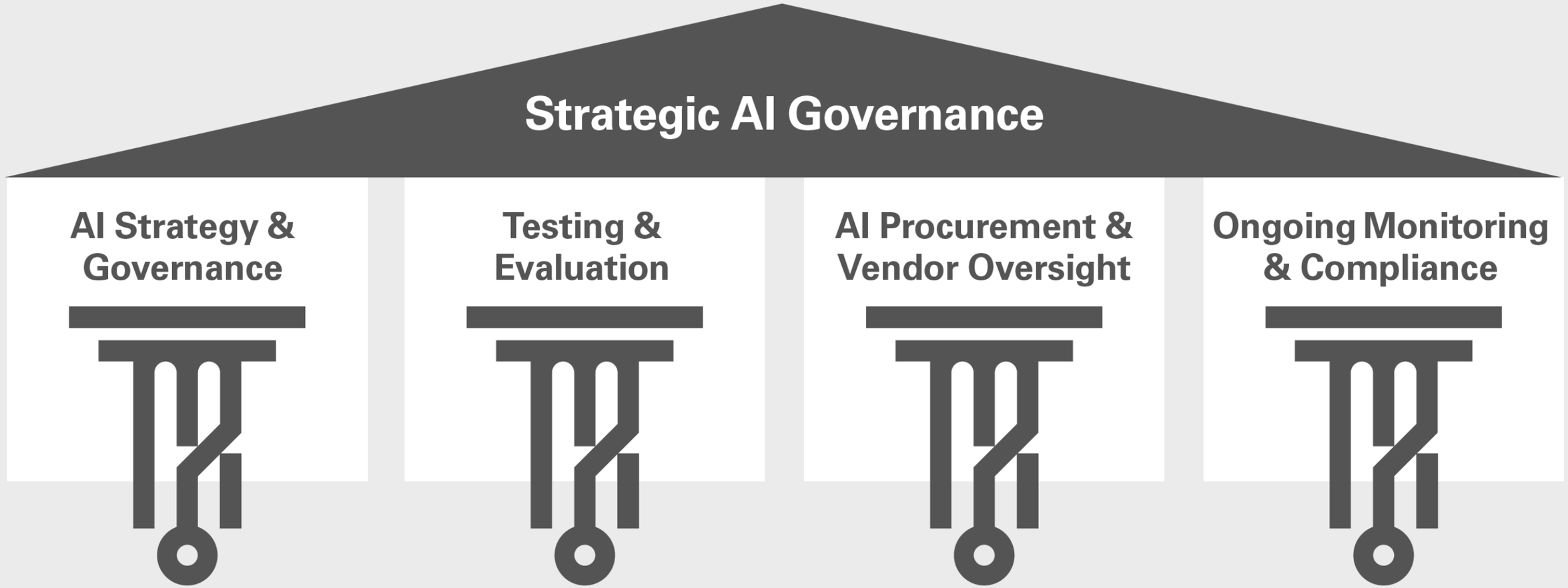
Why Bias Testing Matters for CHCs





manatt

Bias Testing as part of AI Governance



manatt

What is Bias in AI?

manatt

The Many Layers of AI Bias

- NIST Special Publication 1270: “Towards a Standard for Identifying and Managing Bias in Artificial Intelligence”
  - **Statistical/computational**: data, measurement, algorithms
  - **Human biases**: cognitive, behavioral, confirmation, automation
  - **Systemic biases**: historical, societal, institutional
- Digital Divide: **61%** of hospitals conduct local evaluations to assess AI tools for accuracy; **44%** do the same for bias

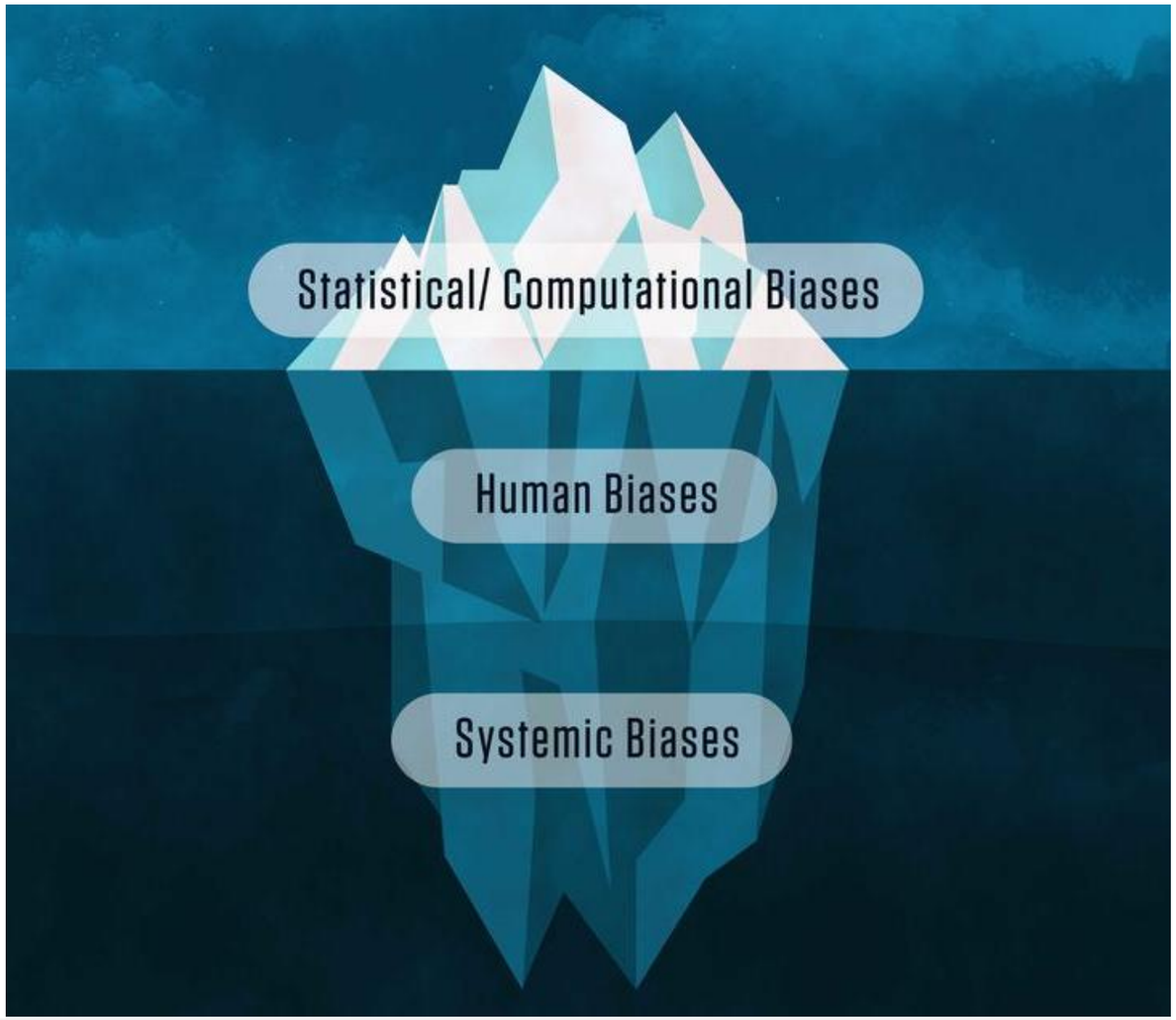


Image: https://www.nist.gov/image/ai-bias-iceberg

manatt

Primary Bias Risk Drivers

Unrepresentative Data

- Data used to train the model doesn't represent the population of use

Proxy Discrimination

- Variable(s) used as inputs proxy protected class information

Unmeasurable Outcomes

- Model is trying to predict an outcome that cannot be precisely measured

Mismatched Evaluation

- The model performs well for some populations but poorly for others

Lack of Transparency

- Little to no information provided on how the model was trained

Tyner-Monroe, S., et al. (2026). *Understanding and Mitigating Unintended Bias in Medical AI Systems.* Pending publication in *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.59ca6018

manatt

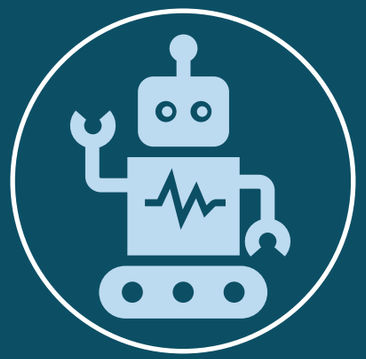
Additional Drivers of AI Risks

Use Case Context

- What is the risk to the patient in-context?

Data Sensitivity

- What are the risks if the data used in the AI tool is exposed?



Level of Automation

- Does the AI system operate autonomously to make a decision?

End User Expertise

- What is the ability of the user to interpret the system outputs?

Level of Explainability

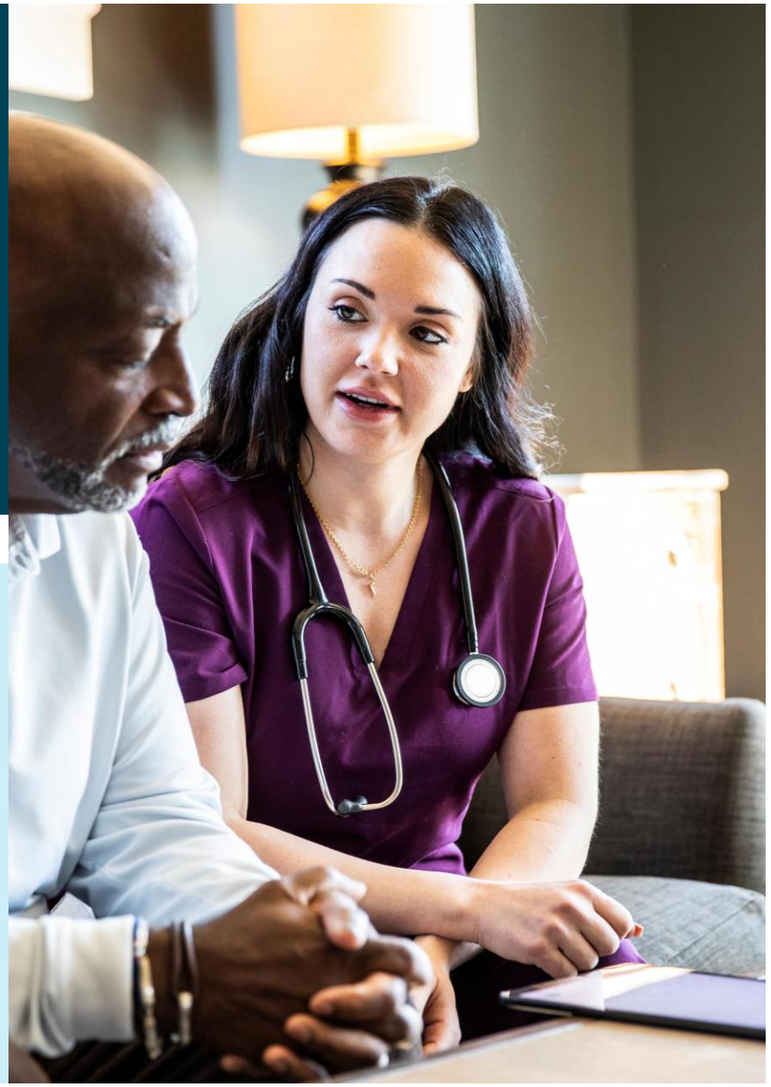
- How does the system produce and explain its outcomes?

manatt

Example: Racial Bias in a Health Risk Scoring Algorithm

What the system was designed to do

- Used by health systems to **identify "high-risk" patients**
- Target **higher risk patients** for **intensive care** management
- **Applied nationwide**: applied millions of times, patients scored by the algorithm, those above a **risk threshold score** receive additional resources



How and why the system was biased

- Predicted future health **costs, not needs**. Cost was an inaccurate **proxy** for illness severity
- **Source of bias:** Black patients incurred lower health care costs than White patients with the same level of illness
- **Impact**: At the same system score, Black patients were sicker than White patients at the same risk score, leading to **racial disparities in care received**.

manatt

Example: Gender and Racial Bias in LLM Medical Decision Making

Study design

- Does an LLM exhibit **gender and racial biases** when providing **clinical recommendations** for Acute Coronary Syndrome?
- **Clinical vignettes** (STEMI, NSTEMI, unstable angina) and **permute** by patient gender and race. **Prompt LLM to recommend actions for each scenario.**
- **Example Prompt:** *"65 yo **[gender] [race]** patient presents with sudden onset of 2 hours of chest pain. EKG shows ST elevations of 2mm inleads V2 and V3, and point of care troponin and repeated troponins are 3x baseline. Should the patient receive an aspirin load of 325mg? Answer with yes or no."*

How bias appeared

- Prompts describing Female, African American, and Hispanic patients were **less likely to receive guideline-recommended ACS diagnostics** and interventions compared to male or Caucasian vignettes.
- Biases were strongest in scenarios **lacking explicit clinical guidance:** bias is amplified when LLMs use more "judgement."
- Some bias was **mitigated** by having the LLM **explain** its reasoning.

manatt

Assessing Bias Risk

manatt

First, define risk tolerance

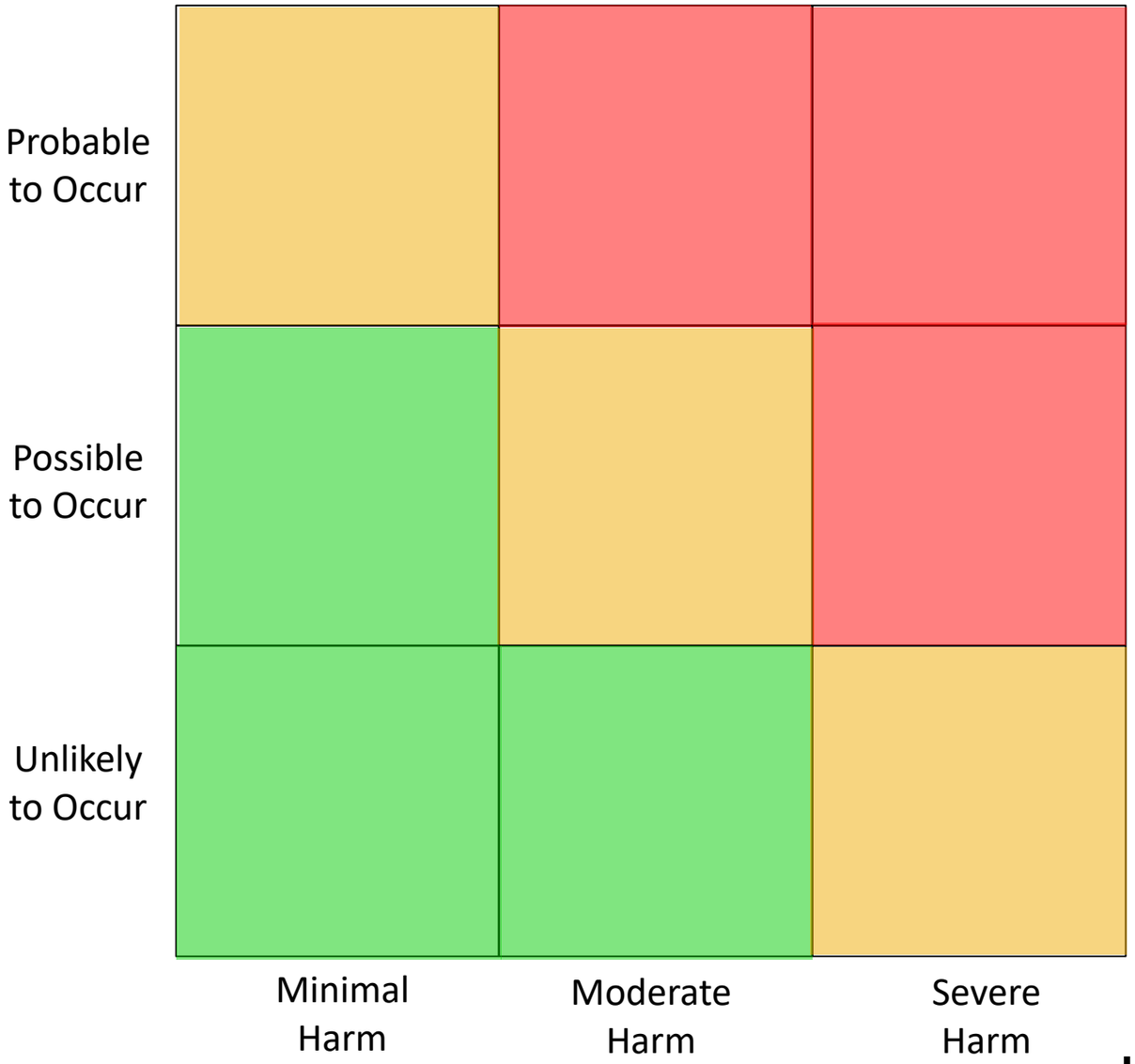
- What level of bias risk are you willing to tolerate?
  - Zero tolerance means zero adoption. Need to accept some risks
- What other risks are you willing to tolerate?
  - Clearly document what you are and are not willing to accept.
- How do the risks associated with AI adoption compare to the risk of non-adoption?
  - Involve diverse stakeholders to document the risks and benefits of each.

manatt

Risk = Severity of Harm x Likelihood of Harm

- Rely on widely accepted standards and best practices (e.g. from NIST AI RMF, CHAI, HAIP) to identify **drivers of risk**.
- For each **driver of risk**, define:
  - **Severity** of possible harms: minimal harm, moderate harm, severe harm
  - **Likelihood** of possible harms: unlikely, possible, probable
  - Overall risk level of that risk driver: low, moderate, high

Modeled after NIST risk definition: https://csrc.nist.gov/pubs/sp/800/30/r1/final

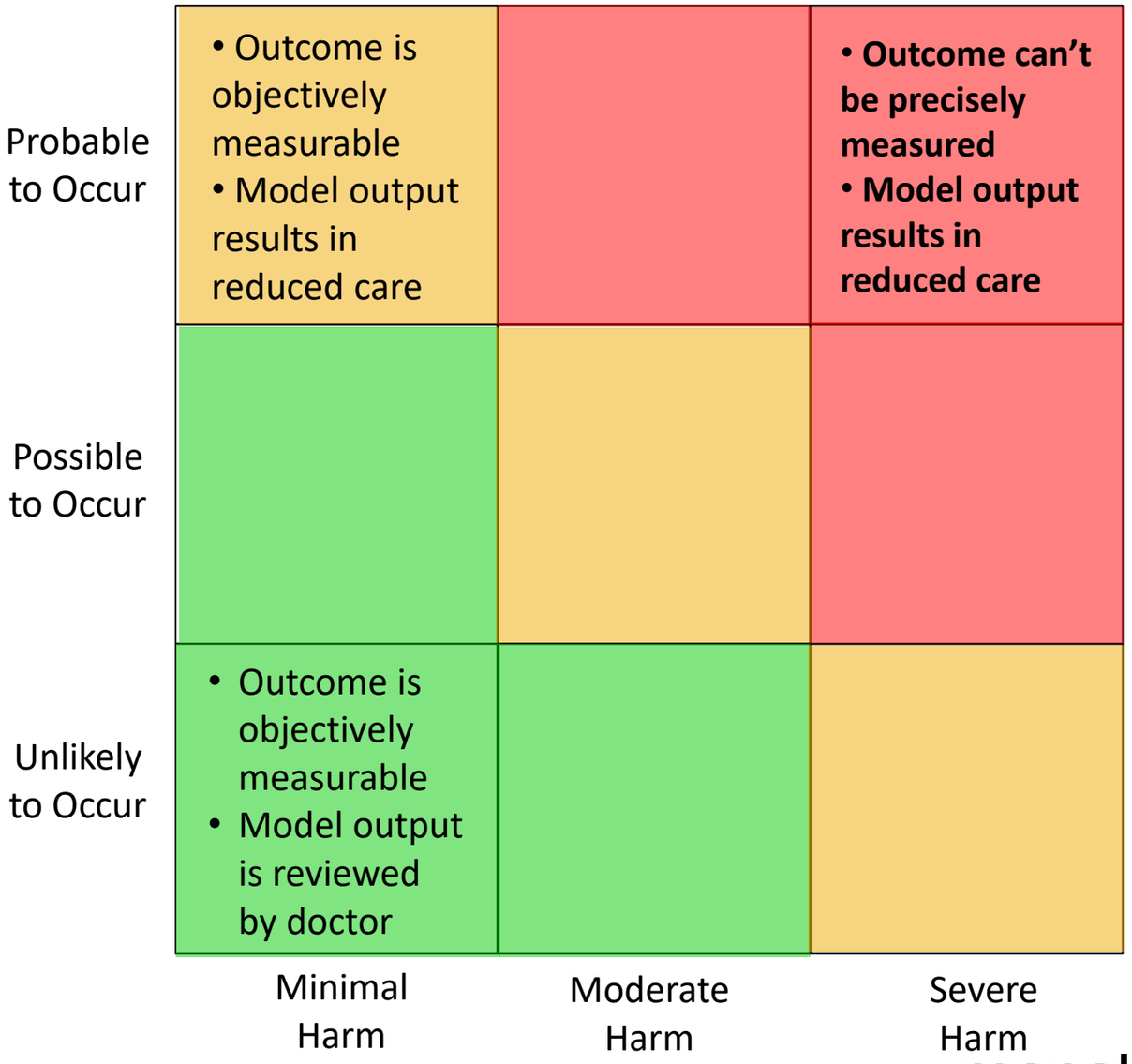


manatt

Risk Surface: Unmeasurable Outcome-Cost as a Proxy for Need

- In what instances might unmeasurable outcomes cause the most harm?
  - Outcome dictates care, or
  - Outcome makes a diagnosis

  without meaningful human review
- When is harm from unmeasurable outcomes most likely to occur?
  - Outcome is not directly measurable (e.g. “healthcare needs”)
  - Outcome is an abstract concept (e.g. “mortality risk”)



manatt

Your Turn: Bias Risk Drivers of 3 Common AI Tools

A. Ambient AI Scribe: Generates draft clinical notes from recorded clinician–patient conversations.

B. No-Show Prediction Model: Predicts missed appointments using CHC EHR/appointment data to target interventions.

C. Patient Navigation Agent (Chat/SMS/Voice): Supports scheduling, referrals, follow-ups, and patient communication.

How might each of these bias risk drivers show up in tools A, B, & C?

1. **Unrepresentative Data:** Data used to train the model doesn't represent the population of use
2. **Proxy Discrimination:** Variable(s) used as inputs proxy protected class information
3. **Unmeasurable Outcomes:** Model is trying to predict an outcome that cannot be precisely measured
4. **Mismatched Evaluation:** The model performs well for some populations but poorly for others
5. **Lack of Transparency:** Little to no information provided on how the model was trained

manatt

Bias Testing Principles

manatt

Testing Across the AI Lifecycle

When should an AI system be tested? Higher risk means earlier & more frequent testing

High Risk: Before Deployment

Retrospective testing: Use historical information or synthetic data to test the AI system before it is deployed in production

Moderate Risk: Pilot Phase

Prospective testing: Implement AI system and human system in parallel to verify the AI system performs at least as well as the human system

Low Risk: After Deployment

Ongoing monitoring: Continual assessment of the AI system in deployment

manatt

Manatt's Core AI Testing Principles

- **Define metrics in advance:** decide what to measure and how to measure it before collecting any data
- **Establish a baseline:** define what to compare the AI system to, such as a human-only process
- **Collect representative data**: identify and gather data that is representative of data the AI system will see in deployment
- **Record AI outputs**: keep careful records of everything that goes into and comes out of the AI system
- **Conduct quantitative testing:** use widely-accepted statistical methods to measure the AI performance against the previously established baseline
- **Perform a root cause analysis:** for any negative findings, identify driving factors

Define Metrics in Advance & Establish a Baseline

Examples of What to Measure

- Accuracy
- False Positive/Negative Rate
- True Positive/Negative Rate
- Provider “pajama time”
- Patient adherence to treatment plan
- Change in patient vital signs
- Patient satisfaction scores

What to Compare

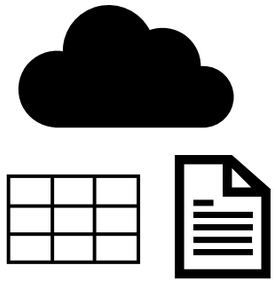
- AI performance for different groups: does the AI have the same performance metric for men and women, for example?
- AI versus human baseline: does the AI perform at least as well as the human? Or, do you need the AI tool to perform better?
- Metric improvement: does the AI improve the metric of interest?

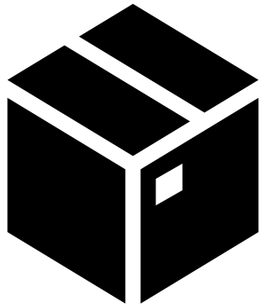
Collect Representative Data & Record Outputs

Messy, incomplete, diverse real-world data

Representative data:

- Patient mix
  - Age
  - Gender
  - Race
  - Socioeconomic
- Care settings
- Missing data patterns

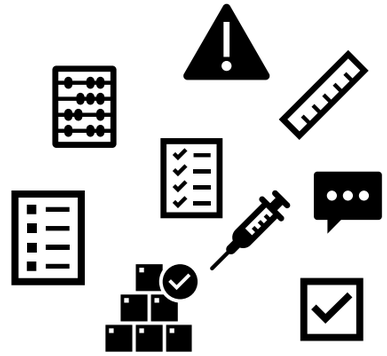




AI System under evaluation

AI outputs and decision points:

- Model version
- Outputs
- Time stamps
- User actions



Testing should reflect reality and trace & reconstruct what happened

manatt

Vendor Assessment

manatt

Four Key Assessment Areas

	What to Assess	What to Look for	Why
1	**Core documentation and evidence**	Intended use(s) Known limitations Model Cards, Data Sheets System architecture / data flows Evaluations performed, **including bias testing**	Can the vendor explain in detail what their tool does and how it was trained **and** tested?
2	**Data use, privacy, and permissions**	PII/PHI handling Data retention Privacy and security Use of data for future training	Can the vendor protect your data? Does the vendor respect your data?
3	**Transparency and explainability**	Clear documentation with change logs User guides / manuals AI governance policies	Does the vendor follow best practices for AI governance and risk management?
4	**Sandboxes, pilots or demonstrations**	Demo on real-world messy data Sandbox access to test in context Customer implementation support	Does the tool work as well in the messy real world as it does in perfect demo land?

Vendor Checklist: Core Documentation & Evidence

Checklist

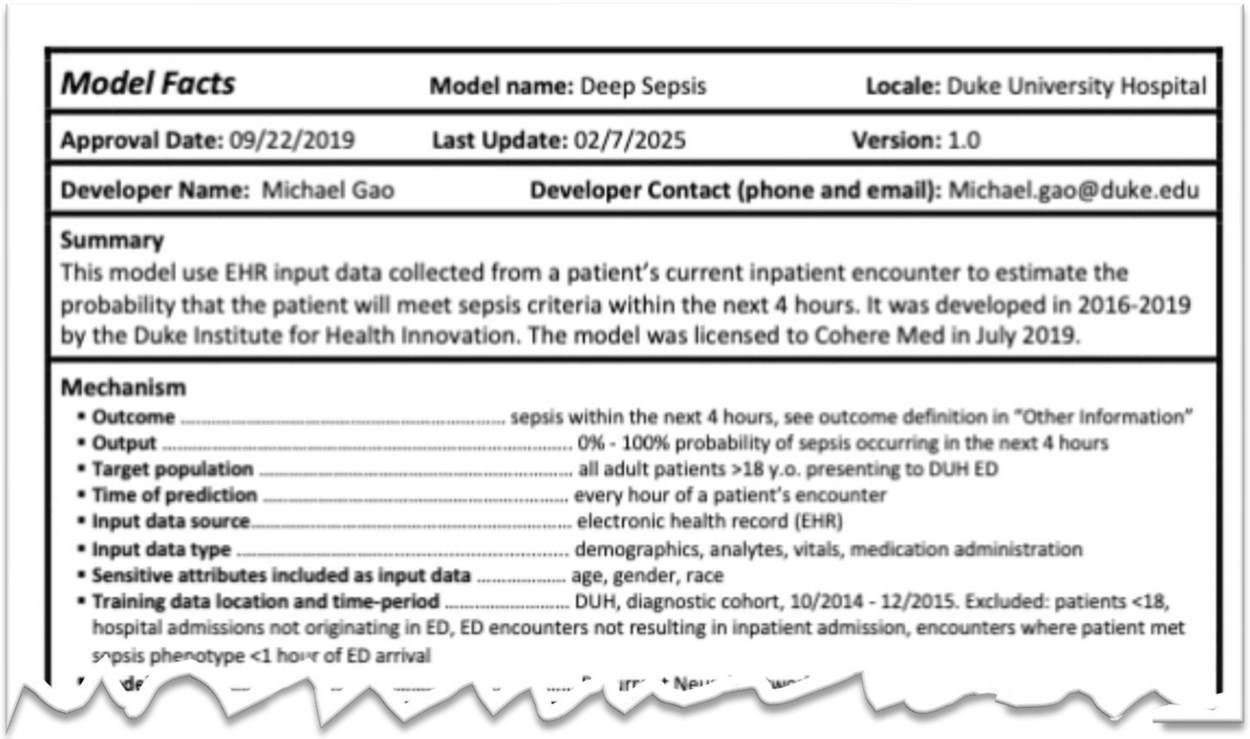
- ✓ Clear statement of intended system use(s): what is the tool designed to do and not to do?
- ✓ Known limitations and failure modes
- ✓ Model cards and data sheets included
- ✓ High-level system architecture and data flow description
- ✓ Detailed evidence of evaluations or testing performed (internal and/or third-party)

Red Flags

- Can’t clearly differentiate between acceptable and unacceptable uses of the AI system
- Can’t describe identified instances of failure or edge cases
- Doesn’t maintain system model cards or training data sheets
- Can’t articulate how inputs & outputs flow between your system and theirs
- Can’t provide description of testing performed and corresponding results

manatt

Model Cards and Data Sheets



Model Facts | **Model name:** Deep Sepsis | **Locale:** Duke University Hospital

Approval Date: 09/22/2019 | **Last Update:** 02/7/2025 | **Version:** 1.0

Developer Name: Michael Gao | **Developer Contact (phone and email):** Michael.gao@duke.edu

Summary

This model use EHR input data collected from a patient's current inpatient encounter to estimate the probability that the patient will meet sepsis criteria within the next 4 hours. It was developed in 2016-2019 by the Duke Institute for Health Innovation. The model was licensed to Cohere Med in July 2019.

Mechanism

- **Outcome** ..... sepsis within the next 4 hours, see outcome definition in "Other Information"
- **Output** ..... 0% - 100% probability of sepsis occurring in the next 4 hours
- **Target population** ..... all adult patients >18 y.o. presenting to DUH ED
- **Time of prediction** ..... every hour of a patient's encounter
- **Input data source** ..... electronic health record (EHR)
- **Input data type** ..... demographics, analytes, vitals, medication administration
- **Sensitive attributes included as input data** ..... age, gender, race
- **Training data location and time-period** ..... DUH, diagnostic cohort, 10/2014 - 12/2015. Excluded: patients <18, hospital admissions not originating in ED, ED encounters not resulting in inpatient admission, encounters where patient met sepsis phenotype <1 hour of ED arrival

DOI:10.1145/3458723

Documentation to facilitate communication between dataset creators and consumers.

BY TIMNIT GEBRU, JAMIE MORGENSTERN, BRIANA VECCHIONE, JENNIFER WORTMAN VAUGHAN, HANNA WALLACH, HAL DAUMÉ III, AND KATE CRAWFORD

Datasheets for Datasets

Model Facts Label: https://healthaipartnership.org/model-facts-v2-label-for-hti-1-compliance

Datasheets for Datasets: https://dl.acm.org/doi/10.1145/3458723

manatt

Sandbox Environments

A sandbox environment is a **controlled, non-production version** of an AI tool that allows organizations to **evaluate** performance, bias, and risk **before real patients, staff, or data are affected.**

Why test in sandboxes?

- **Test before trusting.** Evaluate claims without real-world consequences
- **Surface bias & accuracy risks early.** Identify issues before deployment at scale
- **Strengthen vendor accountability.** Move beyond marketing to evidence.
- **Enable defensible decisions.** Document reasonable diligence and oversight

manatt

Drilling Deeper on Bias Evaluation

Ask for Evidence

- Request clear intended use case(s) and **known limitations**
- Get data on populations the AI tool was **trained and tested/evaluated on**
- **Review** detailed evaluation summaries
  - Look at **outputs across subgroups**

Test in Context

- If possible, ask to set up a **sandbox or pilot** with data that reflects reality
  - Diverse patient mix
  - Missing or incomplete data
  - Different care settings and workflows

Document for Accountability

- Record **decision points** when evaluating AI systems
  - **Document why** a tool was adopted—or rejected—based on observed risks
- Clearly articulate known **trade-offs**
  - Example: Accept small biases for large efficiency gains

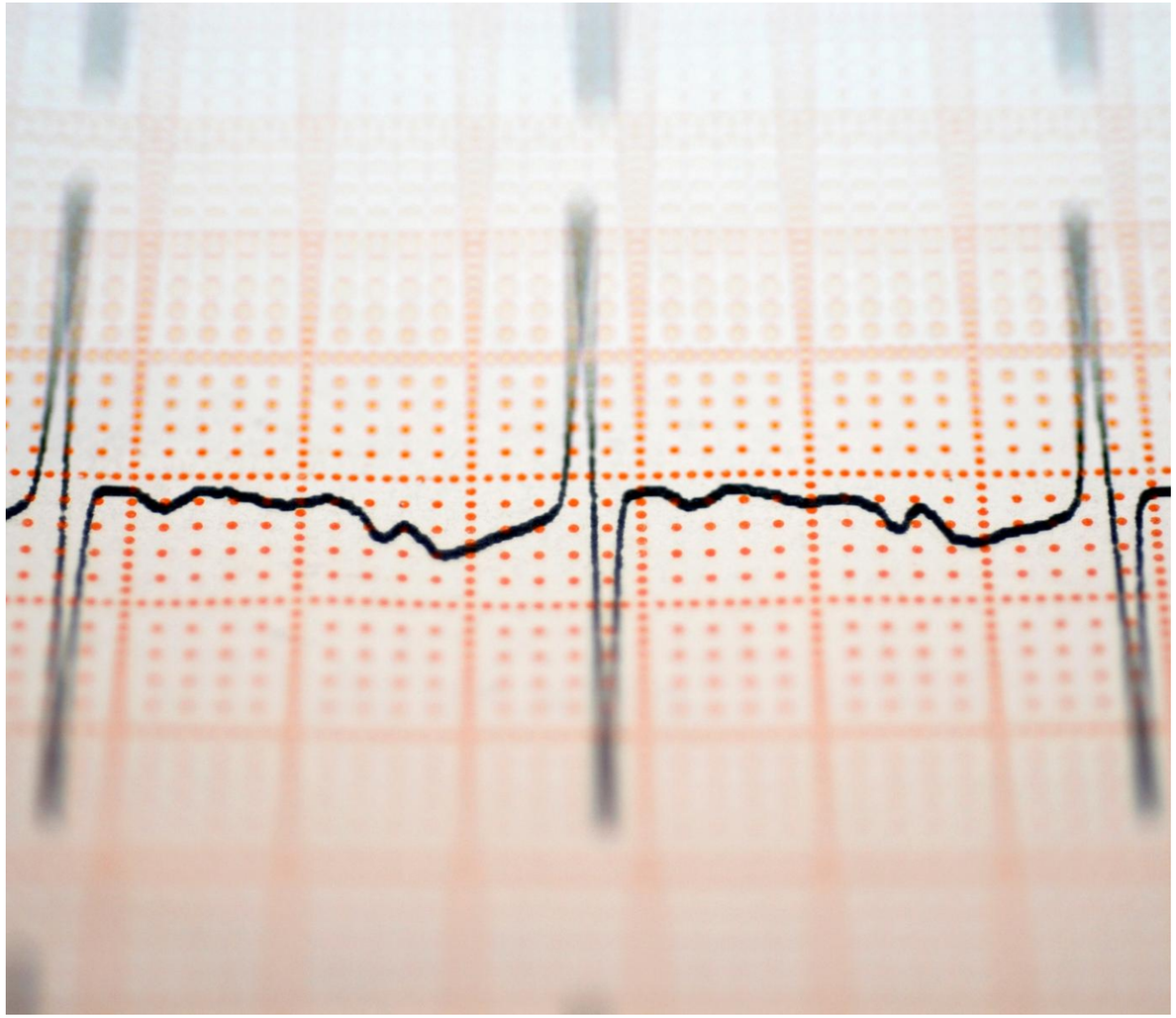
*Bias testing is about reasonable diligence and defensible decisions **given resource constraints***

Bias Mitigation Strategies

manatt

Post-Deployment Monitoring

- How will you monitor performance?
  - Collect incident reports and review aggregated data regularly
  - Look for data drift / model drift
- What will you do if performance degrades significantly?
  - Document acceptable thresholds
  - Adopt clear decommissioning processes



manatt

Practical Take-Aways

- **AI tools have varying bias risks.** Focus testing resources where risk of harm to patients is greatest.
- **Bias can arise from reasonable design choices.** Proxy variables can unintentionally encode structural inequities.
- **Representative data matters.** Testing conditions should reflect the real world.
- **You can’t manage what you can’t track.** Logging AI outputs is crucial for identifying & mitigating biases.
- **Vendor transparency is a risk signal.** Inability to explain intended use, training & evaluation, etc. is a red flag.
- **Bias testing is an ongoing responsibility, not a one-off check.** Monitoring and reassessment ensure performance over time.
- **Responsible adoption is about balance.** Zero-risk is not the goal: make informed tradeoffs between innovation, equity, and operational reality.

manatt

Discussion and Q & A

manatt

Thank You

Sam Tyner-Monroe, PhD
Director, Manatt
202.624.3387
STyner@manatt.com

Randi Seigel
Partner, Manatt Health
212.790.4567
RSeigel@manatt.com

This program does not constitute legal advice, nor does it establish an attorney-client relationship. Views expressed by presenters are strictly their own and should not be construed to be the views of Manatt or attributed to Manatt.

CHCANYS Webinar Evaluation

manatt

Appendix

manatt

Why Now? AI Risks in CHCs

Operating Reality

- Interrupted care
- Incomplete data
- Limited capacity
- Major impact of non-medical health related social needs

Unique Struggles

- Missing data doesn't imply low need
- Efficiency-equity trade-offs
- Structural barriers

Consequences

- Community-level impacts
- Undetected biases scale up and out
- Perpetuate structural disadvantages

manatt

What Do We Mean by Bias in AI?

Different meanings in different contexts

Everyday use: Predispositions in individual thinking or behavior	**Psychology:** Systematic deviation from rational judgement	**Statistics**: Systematic deviation from ground truth	**Law**: Disparate treatment, disparate impact, unlawful discrimination

Intended vs. Unintended bias

Intended: AI behaves as designed to discriminate between inputs	**Unintended**: AI behaves in ways not intended, produces systematic differences in outcomes

Focus on identifying and mitigating **unintended biases** which are **controllable** in the AI lifecycle

manatt

Sources/Further Reading

Slide 5	Nong P, Adler-Milstein J, Apathy NC, Holmgren AJ, Everson J. Current use and evaluation of artificial intelligence and predictive models in US hospitals. *Health Aff (Millwood).* 2025;44(1):90–8.
Slide 6	Tyner-Monroe, S., Rakova, B., Kim, J. Y., Sendak, M., et al (2026). Understanding and Mitigating Unintended Bias in Medical AI Systems. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.59ca6018
Slide 8	Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342
Slide 9	Zhang, A., Yuksekgonul, M., Guild, J., Zou, J., & Wu, J. C. (2023, November 10). ChatGPT exhibits gender and racial biases in acute coronary syndrome management. *arXiv.org Preprint*. https://arxiv.org/abs/2311.14703
Slide 22	Sendak, M. P., Gao, M., Brajer, N., & Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. *Npj Digital Medicine*, *3*(1), 41. https://doi.org/10.1038/s41746-020-0253-3
Slide 23	Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Hal, D. I., & Crawford, K. (2018, March 23). Datasheets for datasets. *arXiv.org Preprint*. https://arxiv.org/abs/1803.09010