# PACT-CARE™: A Human-in-the-Loop Framework for Responsible and Useful AI in Healthcare

Author: Sudhir Shandilya | NextBrightPath



## Executive Summary

Artificial Intelligence (AI) in healthcare often boasts high technical accuracy in controlled settings, yet many models fail to deliver value in real-world care. The problem is not simply model performance, but *usefulness* – integration into clinical workflow, trust by end-users, and tangible patient benefit. To bridge this gap, we propose **PACT-CARE™**, a universal human-in-the-loop (HITL) framework tailored to healthcare. This 8-step loop (Patient & Problem → Action Policy → Capacity & Context → Thresholds & Trade-offs → Compliance & Regulation → Adoption & User Experience → Reliability & Recalibration → Equity, Evidence & Economics) embeds human oversight at every stage, ensuring AI systems are *accountable, trusted, and safely integrated*. PACT-CARE™ synthesizes best practices from existing guidelines – including the FUTURE-AI principles for trustworthy AI, Stanford HAI's "usefulness lens" (which emphasizes net benefit over mere accuracy), the Stanford **FURM** evaluation framework (Fair, Useful, Reliable, Measurable), and global regulatory standards (FDA's Good Machine Learning Practice, ONC's algorithm transparency requirements, WHO's AI ethics guidance, and the EU AI Act).

We present a **toolkit** comprising a PACT-CARE™ scorecard, project canvas, and transparency datasheet to operationalize these principles. Three use cases illustrate step-by-step application in clinical workflows (imaging triage for pneumothorax, preventive care for cardiovascular risk, and operational no-show prediction), with concrete measures and outcomes. These case studies demonstrate how PACT-CARE™ helps move healthcare AI from mere accuracy to real-world usefulness – delivering measurable improvements in care delivery, safety, equity, and trust.

## Introduction: Beyond Accuracy

Despite an abundance of AI models in healthcare, most remain stuck in "pilot-itis" – successful in pilots or papers but not widely adopted in practice. Why does adoption lag? A key reason is that AI systems have traditionally been evaluated on technical metrics like AUROC or sensitivity, rather than on *usefulness* metrics that matter in clinical care (e.g. net patient benefit, clinician workload impact, cost-effectiveness). Many highly accurate models don't translate into value at the bedside.

The **Epic Sepsis Model** exemplifies this gap. It was implemented in hundreds of hospitals, yet independent studies found it had poor positive predictive value (~7.6% in one external validation)[1]. In practice it generated frequent false alarms that **overwhelmed clinicians** with alerts for patients who never developed sepsis[1][2]. By contrast, simpler tools like the ACC/AHA **ASCVD Risk Calculator** (for 10-year cardiac risk) have seen broad clinical adoption. The risk calculator succeeds because it is **interpretable**, anchored in guidelines, and embedded in workflow with clear human oversight (the physician reviews the calculated risk with the patient and decides on statin therapy)[3][4]. In other words, it was designed for *usefulness* and trust from the start. Studies show that providing clinicians with patient-specific reminders based on the ASCVD calculator can significantly increase appropriate statin prescriptions – e.g. a randomized trial reported a ~10% absolute rise in high-intensity statin use among intervention patients[3], roughly one additional patient treated per 10 reminders sent[4].

The lesson: **accuracy is not enough**. A model must integrate into decision-making in a net beneficial way[5][6]. It should trigger actions that improve outcomes without unduly burdening staff or causing harm. Ultimately, *AI must be useful, trustworthy, and supervised by humans* to succeed in healthcare[7][8]. PACT-CARE™ is our proposed framework to ensure these criteria are met.

## The Foundations of PACT-CARE™

PACT-CARE™ is built on a convergence of best-practice principles from academia, industry, and regulators:

- **FUTURE-AI (Trustworthy AI Principles):** An international consensus framework outlining six key principles for medical AI – *Fairness, Universality, Traceability, Usability, Robustness,* and *Explainability*[9][10]. These ensure AI tools are unbiased, generalizable, transparent in development, user-friendly, resilient to perturbations, and able to explain their reasoning.

- **Stanford HAI's Usefulness Lens:** A paradigm shift championed by Stanford's experts that emphasizes evaluating *clinical usefulness* over standalone model accuracy[5]. Nigam Shah and colleagues argue we must ask *"Is the model actionable and beneficial in practice?"* rather than obsessing over small differences in AUC[5]. This entails analyzing whether the healthcare system has the **capacity** to act on the model's predictions and whether those actions truly help patients in a cost-effective way[11][6]. In short, *a model's worth is measured by outcomes, not just outputs*[6].

- **FURM Framework (Fair, Useful, Reliable, Measurable):** A practical evaluation approach developed at Stanford Health Care to decide if an AI model is ready for deployment[12]. FURM involves an ethical review for fairness (identifying any value misalignment or bias), simulations to estimate *usefulness* in workflow, tests for *reliability* and robustness, and definitions of *measurable* impact (financial

projections, clinical benefit)[12]. This corresponds to asking **What and Why** (Fairness, Utility), **How** (Reliability of the model and workflow integration), and **Impact** (Measurable outcomes and improvement). In published assessments using FURM, only 2 of 6 AI solutions reviewed were deemed ready to implement – highlighting the rigor needed to filter hype from reality[13][14].

- **Global Regulatory Anchors:** PACT-CARE™ aligns with emerging regulations and guidelines for AI in healthcare:

- *FDA (United States):* The FDA's 10 guiding principles of **Good Machine Learning Practice (GMLP)** stress life-cycle oversight of AI/ML medical devices. For example, Principle 7 focuses on the **performance of the human-AI team** (i.e. how well the tool improves decisions with a human in the loop)[15], and Principle 10 mandates continuous monitoring of deployed models with managed re-training to ensure safety[15]. The FDA is also pioneering the concept of a **Predetermined Change Control Plan (PCCP)** for "learning" AI devices – essentially a regulatory-approved plan for how an AI model can update itself over time. In late 2024, FDA (with Canada's and the UK's regulators) released guiding principles for PCCPs[16][17], emphasizing that manufacturers pre-specify what aspects of an algorithm may change, how they will be controlled, and how any modifications' impact will be assessed[17]. This approach aims to permit safe ongoing improvement of AI software without compromising patient safety or requiring constant re-approval.
- *ONC (United States):* The U.S. Office of the National Coordinator's **HTI-1** final rule (2024) introduced new certification criteria for *Decision Support Interventions (DSIs)* in electronic health record systems. Any predictive algorithm used within certified health IT must provide **algorithm transparency** via "source attributes" – plain-language disclosures about the model's input data, logic, performance, and intended use[18][19]. The goal is to let end-users evaluate if a given AI tool is *"fair, appropriate, valid, effective, and safe (FAVES)"*[18]. For example, a sepsis risk score DSI would need to reveal its data sources, how it was validated, its sensitivity/specificity in relevant populations, and whether a human is expected to review the prediction. This transparency empowers clinicians to know what's under the hood of AI recommendations they encounter.
- *WHO (Global):* The World Health Organization has strongly advocated for ethics and human rights in health AI. Its **2021 AI Ethics Guidance** articulated six core principles – including *protecting human autonomy, promoting safety, ensuring transparency, fostering accountability, ensuring equity,* and promoting sustainable AI[8]. Notably, the first principle is that **humans should remain in control of healthcare decisions**: AI must not replace clinician judgement or patient consent, and clinicians should be able to override AI recommendations[20]. The WHO also stresses continuous risk monitoring and human safety checks – *"developers should continuously monitor AI tools to make sure they're not causing harm"*[21]. In 2024, WHO released specific guidance on the rise of **Large Language/Multimodal Models (e.g. GPT)** in health care. It warns of risks like false or biased outputs and

**"automation bias"**, where clinicians or patients might over-rely on AI recommendations[22][23]. WHO calls for rigorous oversight and *human-in-the-loop safeguards* when deploying generative AI in clinical settings[23][24].

- *EU (Europe):* The European Union is enforcing comprehensive regulation via both the existing Medical Device Regulation (MDR/IVDR) and the upcoming **AI Act**. Many AI clinical decision tools qualify as **Software as a Medical Device (SaMD)** and thus already must meet MDR requirements for safety, performance, and risk management. In addition, the EU AI Act – which entered into force in 2024 – classifies most healthcare AI systems as "high-risk," subjecting them to additional requirements such as strict risk assessment, data governance to prevent bias, transparency to users, and **human oversight mechanisms** to minimize automation harm[25][26]. The Act explicitly requires that high-risk AI be designed so that human operators *"can oversee the system's functioning and intervene or disable it if necessary"*. To support implementation, the European Medicines Agency (EMA) issued a 2024 **Reflection Paper on AI in medicine** aligning with the AI Act's principles. EMA emphasizes a *risk-based approach* and states that if an AI's logic is a "black box," sponsors must at least demonstrate **interpretability with human oversight** in its use – meaning clinicians can obtain explanations or have a process to validate the AI's outputs[27]. In short, the EU is mandating a *safety net of human control and accountability* around medical AI tools.

These foundations all point to the same conclusion: effective healthcare AI requires **more than algorithms**. It requires *socio-technical integration* – fairness and usability by design, alignment with clinical capacity and policy, continuous monitoring, and above all, keeping *humans in the loop*. PACT-CARE™ brings these pieces together into a single actionable framework.

## The PACT-CARE™ Framework (with Human-in-the-Loop)

PACT-CARE™ is organized as an **8-step iterative loop** covering the entire lifecycle of an AI solution in healthcare. At each step, a *human-in-the-loop checkpoint* ensures the system remains aligned with clinical goals and ethical standards. Below we outline each component of PACT-CARE™, including its purpose and example metrics:

- **P – Patient & Problem:** Identify a specific *patient-centered problem* to solve. Define the outcome or decision that the AI is meant to improve, and quantify the baseline gaps in care. Importantly, engage clinicians (and patients if possible) at the scoping stage – the people feeling the pain point should help define the problem. *HITL role:* domain experts validate that the problem is real and the target outcome is meaningful. **Example:** In emergency radiology, the problem might be delays in treating pneumothorax. Baseline analysis might show a *"median 55 minutes from X-ray to chest tube placement"* for pneumothorax patients – indicating a delay opportunity to improve. This metric gives a patient-centered goal (reduce time to treatment). Clinicians confirm that faster intervention would indeed benefit patients (e.g. reduce distress or prevent deterioration). If not, the project should be re-

scoped. By starting with a clear problem definition, we avoid deploying AI "because we can" and focus on *what truly needs fixing*.

- **A – Action Policy:** Ensure that **every AI prediction is mapped to a specific human action in workflow**. In healthcare, an algorithm on its own does nothing – value comes only when it *triggers an appropriate intervention*. Define in advance *who* will act on the AI output, *what* they will do, and *how* this fits into clinical processes. *HITL role:* a designated human decision-maker (clinician or operator) **reviews each AI output and has authority to override or modify the suggested action** before it affects the patient. This is critical for safety and accountability[20]. **Example:** In an AI-driven chest X-ray triage system, the action policy might be: "If the AI flags a possible pneumothorax on an ER patient's X-ray, then the on-call radiologist immediately reviews that image **ahead of others** and, if confirmed, activates a STAT alert to the emergency team." Here the radiologist is the human-in-loop gatekeeper – the AI doesn't directly call the code or page the surgeon; a doctor verifies the finding[28]. By contrast, if an AI's output doesn't clearly map to an action, it's a red flag (e.g. a risk score that doesn't come with any treatment protocol or recommendation is unlikely to be useful). An explicit action policy also helps with regulatory clarity: if the AI is merely *advice* for a human (Decision Support), versus automating a decision (which could make it a medical device), different oversight applies. PACT-CARE™ requires that an identified person (or role) is responsible for acting on and overseeing the AI's output at all times.

- **C – Capacity & Context:** Before deployment, verify that the healthcare system has the **capacity (staff, time, resources)** to *actually perform the actions* the AI will trigger, including the human-in-loop review workload. An AI alert is useless if it recommends an action that can't be taken due to operational constraints. This step evaluates the context of care and whether the AI fits within it. Define an **Action Capacity Ratio (ACR)** – the expected frequency of AI alerts or tasks compared to the available human capacity to handle them. For safe use, aim for ACR ≤ 1 (i.e. do not generate more interventions than staff can reasonably absorb). *HITL role:* workflow managers or front-line staff confirm that the alert frequency is sustainable and adjust thresholds or scope if necessary. **Example:** If an AI system predicts no-shows and suggests outreach calls to patients, but your clinic only has one staffer who can make ~40 calls a day, the AI should not flag 100 patients every day. If initial modeling shows 100 high-risk no-shows daily, you might narrow the criteria (higher risk cutoff) so that perhaps 30–40 patients are flagged – aligning with capacity. We can quantify this as an **ACR = alerts ÷ staff capacity**. In a radiology context, suppose an AI triage for critical X-rays issues 6 alerts per hour and each alert requires ~5 minutes for a radiologist to review and act. That's 30 minutes of extra work per hour – an ACR of 0.5 (50% of available time), which might be acceptable. If ACR were >1 (more alert work than time), the system would *overwhelm staff* and likely be ignored. Many AI projects fail because they don't consider this: for instance, one sepsis alert system inundated clinicians with far more alerts than

they could manage, leading to alert fatigue[2]. PACT-CARE™ explicitly guards against that by aligning AI output to human capacity.

- **T – Thresholds & Trade-offs:** Determine the **operating threshold(s)** for the AI – balancing sensitivity vs. precision – based on explicit *clinical trade-offs and decision curve analysis*. Rather than maximizing technical metrics, choose a threshold that yields the best *net benefit* in practice[29][30]. Decision Curve Analysis (DCA) is a useful method here: it calculates net clinical benefit across a range of threshold probabilities by weighing true positives against false positives, given a certain "harm-to-benefit" trade-off[31][32]. Incorporate the *human review burden* into these trade-offs as well – e.g. a false positive doesn't just cause potential patient anxiety, it also consumes reviewer time. Define metrics like **NNA (Number Needed to Alert)** – how many alerts does the system gives to yield one truly beneficial intervention. If the NNA is too high, the system may do more harm (through distraction and extra work) than good. *HITL role:* clinical leadership participates in selecting the threshold that they feel comfortable with and that aligns with patient risk preferences. **Example:** For a predictive model to recommend statin therapy, a risk threshold of, say, 7.5% 10-year ASCVD risk (per guidelines) might be the trigger for "AI suggests statin." Using DCA on your population, you might test thresholds from 5% to 20%. If threshold is too low (5%), you'd alert on many patients, burdening clinicians and possibly over-treating; too high (20%) and you miss many who could benefit. The analysis might show that around 10% risk yields the highest net benefit (treating those likely to benefit and few unnecessary treatments). Thus you set the AI to flag patients above 10% risk. Additionally, suppose at that threshold the model's PPV is 30%, meaning ~1 in 3 flagged actually have the outcome or benefit – that might be acceptable. In contrast, the earlier Epic sepsis model example effectively had an **alert PPV under 8%**, meaning over 12 false alarms per true case[1]. Such a low yield overwhelmed clinicians and had *negative* net value. PACT-CARE™ emphasizes using domain-informed thresholds (potentially via pilot studies) to balance these trade-offs. We also calculate NNA: e.g. if at 10% threshold, 50 patients are flagged per actual additional statin start achieved, NNA = 50 – is that acceptable? It might be borderline, prompting refinement of the model or threshold. By explicitly managing thresholds and trade-offs, we aim for an AI that *meaningfully improves decision-making* rather than one that is tuned only for maximal sensitivity or AUC.

- **C – Compliance & Regulation:** Map out the **regulatory and compliance requirements** relevant to the AI solution *before* deployment. Depending on the nature of the AI and its use, different rules apply – and all require documentation of appropriate human oversight. *HITL role:* compliance officers or regulatory experts review the plan to ensure it meets standards, and clinicians are assigned to maintain oversight in line with those rules. **Example:** Ask: Is this AI a **medical device** influencing patient care (likely SaMD requiring FDA clearance or CE marking under MDR)? Or is it a **clinical decision support tool** that falls under FDA's CDS

exemption but now is subject to ONC's DSI transparency requirements? If it's a sepsis alert integrated in an EHR, one might classify it as a Clinical Decision Support Intervention – meaning it must provide users with the logic or basis of the recommendation (per ONC's rule) and likely does *not* make autonomous decisions (to stay FDA-exempt). In that case, you must create a **Transparency Sheet** listing things like the model's intended use, input features, performance, and the fact that *a clinician will always review the alert* (this fulfills ONC's transparency and user oversight expectation[18][19]). Alternatively, if it's a stand-alone AI diagnosis app sold to providers (thus likely a regulated device), you need to implement **Good Machine Learning Practices (GMLP)** – e.g. documenting data representativeness, software validation, etc[33] – and perhaps have a Predetermined Change Control Plan if the model will update over time[16][17]. You would define how the human overseers (clinicians) will monitor the AI's performance in real use and how any potential model changes are managed. Compliance also includes aligning with privacy laws (HIPAA) and ethical standards. The PACT-CARE™ framework makes regulatory compliance less daunting by baking oversight and documentation into each step. By the time you reach this "C", you have a clear record of your model's purpose, human controls, performance characteristics, etc., which regulators and accreditors increasingly want to see. (For example, if auditors ask how your AI triage tool ensures patient safety, you can show that every alert is confirmed by a radiologist and that you monitor override rates as a quality check.)

- **A – Adoption & User Experience:** Plan for **workflow integration, training, and user experience (UX)** to drive adoption. Even a highly accurate, well-intentioned AI will fail if end-users find it hard to use or disruptive. This step focuses on delivering the AI's output in a *usable, clinician-friendly manner* and tracking whether it's actually being used as intended. *HITL role:* end-users (doctors, nurses, scheduling staff, etc.) are involved in UI/UX design and pilot simulations. They remain "in the loop" by having easy ways to override or provide feedback on AI outputs in real time. **Example:** If deploying an AI alert, integrate it into the existing systems clinicians already use (e.g., the EHR or PACS) rather than a separate dashboard that requires extra logins. Time the alerts at points in the workflow where they make sense – for instance, a no-show risk alert should surface a day or two **before** the appointment when staff can still intervene, not at the appointment time. Provide a **one-click option to override or dismiss** an AI suggestion with a required reason code (for example, a doctor can click "dismiss – AI incorrect" or "dismiss – patient not eligible" for an alert). This not only respects clinician authority but also *collects valuable data on why the AI was not followed*. Measure **adoption rate** (what percentage of AI recommendations are acted upon vs. ignored) and **override rate** (how often humans reject the AI). If doctors are overriding 80% of alerts, that signals a problem – either the model or the UX (or both) need improvement. Conversely, high adoption with low override (and good outcomes) indicates success. A real-world example: an AI-driven statin initiation suggestion delivered via the EHR during a primary care visit might be adopted, say, 60% of the time initially. If we see an

override reason frequently like “patient already on therapy” or “clinician disagrees with risk,” we can refine the logic or provide better information to the user. By monitoring these metrics, PACT-CARE™ ensures the AI tool is actually *helping* rather than annoying. Remember, **usability is a key principle in trust** – clinicians need to feel the tool fits naturally into their work with clear benefits. Early studies have shown that when AI tools are convenient and transparent, clinicians are much more likely to trust and use them[34][35].

- **R – Reliability & Recalibration:** Once the AI is deployed, continuously monitor its **real-world performance and drift** over time, and be ready to recalibrate or retrain it as needed. No AI model is static – data distributions change (e.g., new strains of a virus can alter sepsis presentation, or patient demographics shift), and model performance can degrade. Also, monitor the **HITL decisions**: if humans are frequently overriding in certain scenarios, investigate why. *HITL role:* a human (or team) is assigned to regularly review model outputs, outcomes, and override logs, and to trigger model updates or escalations when reliability issues arise. **Example:** Track metrics like model sensitivity, false positive rate, and override rate on a rolling basis. If a deterioration is observed – say a sepsis model’s sensitivity dropped from 80% to 65% this quarter – treat it as you would a drop in laboratory quality: root cause analysis and remediation. In fact, a study in *Nature Medicine* observed a sepsis prediction model’s performance **fell 17% within months of deployment** due to changes in patient population and care processes[36]. Such drift underscores why active monitoring is essential. Establish a schedule (e.g. quarterly or monthly) to *recalibrate* the model with more recent data or to retrain it entirely if needed (following your PCCP or update plan if regulated)[16]. Also use the human feedback: if radiologists override the pneumothorax alert mainly for apical blebs mistaken as PTX, update the algorithm to reduce that error. If primary care docs frequently override the statin AI for patients with certain comorbidities, maybe those conditions should be incorporated into the model or exclusion rules. Essentially, the human-in-loop is not just a safety check, but a source of **ground truth feedback** to improve the AI. PACT-CARE™ requires that organizations treat AI models as learning tools that need ongoing quality assurance – akin to how one would regularly service medical equipment or update clinical guidelines. Deployed models should also have a **fallback procedure**: if something seems off (say, a sudden spike in alerts due to a data interface glitch), clinicians can pause or disable the AI until resolved. This R step ensures the AI continues to be *safe and effective in the long run*, not just on day one[15].

- **E – Equity, Evidence & Economics:** Finally, assess the **broader impacts**: is the AI performing equitably across patient subgroups? Does real-world evidence show improved outcomes? And do the benefits justify the costs (ROI)? This step forces us to look at population-level effects and value for money – critical for scaling an AI solution responsibly. *HITL role:* clinical researchers or quality officers periodically analyze outcomes data, and leadership reviews return on investment and business

cases. **Example:** Evaluate fairness metrics – e.g. is the AI's positive predictive value or error rate similar for different demographic groups? If the pneumothorax AI has 30% PPV in the overall cohort, check it in subgroups (young vs elderly, different ethnicities, etc.) – ensure no group has, say, only 10% PPV (meaning they get lots of false alerts) while another has 40%. We might set an **equity tolerance** like "PPV parity within ±5% across groups" and measure that[37][38]. If disparities are found, adjustments or further training may be needed (perhaps the model underperformed on an under-represented group's data). Also monitor if the *human reviewers* exhibit any bias in overrides – for instance, are clinicians overriding AI suggestions more often for certain patient groups? One real trial demonstrated that a model-based outreach program *reduced no-show rates particularly for Black patients (42% → 36%)*, narrowing the disparity with white patients[39][40]. That's an example of using AI to *improve equity*. We should strive for such outcomes and be alert that the opposite (widening gaps) doesn't occur. In terms of **evidence**, design pilot studies or use existing quality metrics to see if the AI is actually improving the target outcomes. For the pneumothorax case, does implementing the AI + radiologist triage *actually lower the average time to treatment* compared to before? If we aimed for a 10-minute reduction, did we achieve it? Encouragingly, published studies of AI chest X-ray triage have reported substantial workflow improvements – one study found **a 57% reduction in median report turnaround time for routine pneumothorax cases (from 345 min to 148 min)** using an AI-prioritized workflow[41][42]. Another saw overall pneumothorax reporting times drop by ~86 minutes with AI assistance[43][44]. Those are huge gains translating to faster patient care. We should document such evidence in our deployment. Finally, **economics**: calculate the financial impact. This includes the cost of the AI system, implementation, and the HITL overhead (e.g. the radiologist time or extra outreach staff time), versus the savings or revenue gains from improved outcomes or efficiency. For no-show predictions, for instance, reducing missed appointments has clear revenue benefits – one analysis estimated that a mere **5% reduction in no-shows could save a 10-physician clinic over $200,000 per year** in recovered revenue[45]. At a macro scale, no-shows cost the US healthcare system an estimated **$150–300+ billion annually**[46], so there is significant ROI potential in fixing this problem. We should ensure the AI intervention is delivering value that exceeds its cost. Perhaps our no-show program used 1 FTE staff and some software fees (costing $80k/year) but reduced missed visits by 25%, yielding an extra $120k in revenue – that's a positive ROI (1.5x in this scenario, meeting our scorecard goal). If the economics don't pan out, that signals the project might not be sustainable without changes (or perhaps the goal needs to be non-financial, like improved outcomes alone). By explicitly examining equity, evidence, and economics, PACT-CARE™ ensures that an AI that *looks good on paper truly does good in practice*, for all groups of patients and in a cost-conscious manner.

In summary, the PACT-CARE™ loop provides a structured path from conception to deployment, with **continuous human oversight** embedded. It's an iterative cycle – e.g.

after evaluating Equity/Economics, we might loop back to adjust the Problem or Threshold if needed and go through again. This ensures **accountability at each stage** and helps avoid the common pitfalls of health AI (unintended bias, misalignment with workflow, lack of real impact).

## PACT-CARE™ Toolkit

To make this framework practical, we provide several tools:

- **PACT-CARE Canvas:** a one-page template that teams can fill out for a given AI project. It has sections for each of the 8 steps (P through E) with guiding questions. For example, under "Patient & Problem" it asks for the clinical problem statement and baseline metrics; under "Action Policy" it asks who the human decision-maker is and what action will be taken; under "Capacity" it asks for expected alert volume and staff capacity estimates; etc. By completing the canvas, stakeholders create a shared mental model of how the AI will function in the real world. It also serves as documentation of design choices and oversight plans – useful for internal governance and external regulators. (This is akin to a project charter or a Lean canvas but specialized for health AI considerations.)

- **PACT-CARE Scorecard:** a simple scoring system to grade an AI solution on each of the 8 components, to decide if it is ready for deployment. Each element (P, A, C, T, … E) is scored 0, 1, or 2 points based on criteria (0 = not addressed or very weak; 1 = partially addressed; 2 = well addressed with evidence). The maximum score is 16 (or 20 if we give some components extra weight). We propose cut-offs such as: **0–9 = "Pause"** (the project is not ready or fundamentally flawed in design), **10–15 = "Pilot Only"** (proceed with caution in a limited setting), **16+ = "Ready to Scale"**. For instance, if an AI model has no human action plan (score 0 on A) or poor capacity fit (0 on C), it will clearly fall in the Pause range – meaning it should be reworked before any go-live. The scorecard forces an honest evaluation beyond just model accuracy, creating a higher bar for "responsibility." It's inspired by the idea that models should be **Fair, Useful, Reliable, and Measurable** (FURM)[47]– if any of those pillars is missing, deployment should be reconsidered.

- **Operating Point Brief:** a concise report of the chosen operating parameters and expected performance of the AI-HITL system. This includes the threshold probability set for alerts, the sensitivity and positive predictive value at that threshold, the **Number-Needed-to-Alert (NNA)**, the **Action Capacity Ratio (ACR)**, and the net benefit (if calculated via decision analysis) at that point. It also outlines the planned staffing (e.g. "radiologist on-call 24/7 to review alerts") and any *equity check* results (performance across subgroups). Essentially, this is the "fact sheet" of how the AI will run in practice. It's analogous to a medication's dose, efficacy, and side effect profile. For example, an operating brief might say: "Model X will flag ~5% of patients (threshold set at risk score ≥0.7). At this threshold, sensitivity ~85%, PPV ~30%. Number-needed-to-alert ≈ 4 (about 1 in 4 alerts leads to the intended

intervention). Two ICU nurses will receive and triage the alerts (ACR ≈0.5 per nurse). Decision analysis shows a net benefit of 0.05 at this threshold versus treat-all or treat-none. No major performance disparity was found between male and female patients (PPV 32% vs 28%)." Such a brief, when shared with stakeholders, provides transparency and aligns expectations. It could be shared with hospital leadership or even patients if appropriate, to communicate what the AI does and how it's controlled.

- **Transparency & Oversight Sheet:** This is a document that captures all the information needed for compliance and trust, much of which overlaps with the ONC's required *"source attributes"* for predictive DSIs[18][19]. It lists: the intended use and target population; the input data the AI uses; the model's development and validation summary (including limitations); performance metrics (discrimination, calibration, error rates); the **human oversight plan** (who reviews outputs, how overrides work, how often model performance is checked); and any guardrails (e.g. the model will not provide output if certain data are missing or if patient is under 18, etc.). Think of it as an enhanced *"model card"* tailored for clinical use and regulatory scrutiny. This sheet should be maintained and updated over the model's life. If the FDA or any auditor asks "How do clinicians know whether to trust this AI?", you can produce this sheet. In fact, under ONC's rule, certified EHRs will likely surface some of this info to end-users – for instance, an ED physician clicking an AI alert might see a tooltip: *"This algorithm was developed on 50,000 cases, achieved 85% sensitivity in validation, and is intended to assist (not replace) clinical judgement. It is monitored by the quality committee monthly."*[19][48]. Providing such transparency is crucial to building user confidence.

By utilizing the canvas, scorecard, briefs, and transparency sheets, organizations can **operationalize the PACT-CARE™ principles**. These tools make the abstract concepts concrete and ensure that all stakeholders (data scientists, clinicians, administrators, compliance officers) stay on the same page. They are also **living documents** – as the project evolves, the canvas and transparency sheet are updated (e.g. after recalibration, or after expanding to a new hospital unit). This creates a continuous quality loop.

## Use Cases (Step-by-Step)

To illustrate how PACT-CARE™ is applied in real scenarios, we present three use cases from different domains (imaging, preventative care, and operations). Each example walks through the PACT-CARE™ steps, demonstrating the decisions, metrics, and outcomes. These are distilled from real-world projects (with some hypothetical elements for illustration):

### Use Case 1: Pneumothorax Triage in Radiology (Imaging AI)

**Patient & Problem:** Patients with an undetected pneumothorax on chest X-ray risk deteriorating while waiting for treatment. In a busy hospital, routine X-ray turnaround might

be 1-2 hours. Our outcome goal is to **reduce the time from image acquisition to chest tube placement** for large pneumothoraces. Baseline data: median ~55 minutes from X-ray to intervention, and some cases taking over 2 hours if scans wait in queue. Clinicians acknowledge this is a serious issue – e.g. a tension pneumothorax can be fatal if treatment is delayed. We scope the AI to target *ER chest X-rays* for STAT triage if pneumothorax is suspected.

**Action Policy:** We implement an FDA-cleared AI (running on the PACS) that can detect pneumothorax on X-rays. The action policy: **if AI flags a possible moderate/large pneumothorax, the on-duty radiologist immediately gets an alert and prioritizes that X-ray for reading within 5 minutes**. The radiologist then confirms the finding and calls the ER physician to initiate treatment (chest tube) if confirmed. *Human-in-loop:* The radiologist must verify; the AI flag alone never directly triggers a clinical action without physician sign-off. This mirrors how some triage AI are used: they reorder the worklist but do not generate final reports[49][50]. We assign a named attending radiologist each shift as the "AI triage monitor" responsible for responding to these alerts. In case of disagreement (AI says pneumothorax, radiologist thinks no), the radiologist's judgement prevails (they can dismiss the alert as false positive). All alerts and actions are logged.

**Capacity & Context:** Our radiology dept has capacity to handle about **6 AI-triggered STAT reads per hour** (beyond normal workload) before it strains the radiologist. We expect, based on incidence, perhaps 10 pneumothorax alerts per day (≈0.4 per hour). That's easily within capacity (ACR ~0.07). Even in surges, we estimate at most 2 per hour occasionally, which is still manageable (ACR 2/6 = 0.33). The radiologists agree this is fine as long as false alarms aren't excessive. We also consider ED context: the ED must be ready to act faster on these findings. The ER physicians and trauma team are on board – they say a heads-up call 30-60 minutes sooner is valuable and they can mobilize a response team quickly. So the clinical context supports the intervention (we won't be alerting into a void). We will monitor how often radiologists have simultaneous multiple AI alerts – if too often, we might need a second reader or adjust thresholds.

**Thresholds & Trade-offs:** The AI model's sensitivity can be adjusted via its operating threshold for the pneumothorax probability. We perform a decision analysis with radiology and ER input: missing a large pneumothorax is very dangerous, so we want high sensitivity – but too many false positives will distract staff. We target a threshold that yields **sensitivity ~95% for moderate/large pneumothorax**, accepting that some small apical pneumothoraces might be missed (those are less urgent). At this threshold, the vendor reports the model's **PPV is ~30%** (so ~1 in 3 alerts is a true pneumothorax needing action). We calculate an initial **NNA ≈ 3.3** (for every 3-4 alerts, one chest tube intervention is actually needed). Is that acceptable? The radiologists believe yes – scanning 3 false alarms to find 1 real critical case is worthwhile, as reading 4 X-rays is trivial compared to missing a chest trauma. An independent validation in literature showed that an AI pneumothorax triage tool achieved about **97% sensitivity and 60% specificity for clinically significant pneumothorax**[51][52]. That specificity (60%) aligns with a roughly 40% false positive rate, which is similar to our settings. We also define an alert exclusion: the AI will not flag if the

patient is already known to have a chest tube or certain post-surgical cases, to avoid redundant alerts. The radiologists will track how many alerts were "unnecessary" (e.g. AI misidentified a skin fold as pneumothorax). If PPV in practice falls much below 25%, we may tighten the threshold. We also consider **Number Needed to Treat (NNT)**: by expediting these cases, how many alerts translate to improved outcomes? If every true pneumothorax alert presumably gets a chest tube 30 minutes sooner, we think every true alert is essentially an "NNT = 1" for faster care. The trade-off is the extra calls for false alarms (maybe a few unnecessary ED preps). We will watch for any negative feedback from ED on false alarms; if ED is getting many "all clear" after mobilization, we might adjust threshold to improve PPV. Overall, we bias towards sensitivity given the high risk of missing a tension pneumothorax.

**Compliance & Regulation:** The AI software is a cleared **SaMD (Software as a Medical Device)**, so it meets MDR/FDA requirements. We still must follow good radiology practice: we informed our institutional review board and obtained appropriate approvals for using AI in clinical workflow. For ONC's DSI transparency, we document the AI's integration in our certified PACS – including its intended use ("alerts radiologist to possible pneumothorax"), input (frontal chest X-ray), and performance. We maintain a **PACT-CARE transparency sheet** stating that *"a board-certified radiologist remains the final interpreter on all studies; the AI is an assistive triage tool."* This addresses the new US rules to disclose predictive algorithms to users[48][20]. We also train radiologists on what the AI can/can't do (for instance, it might not detect very small apical pneumothoraces – so they shouldn't assume "no alert means no pneumothorax"). From a liability standpoint, the radiologist on duty is still responsible for the reading. We log each alert outcome for QA. All these steps ensure **traceability and accountability**, aligning with guidelines[53]. Finally, we comply with FDA's post-market monitoring advice (Principle 10 of GMLP): we'll monitor performance drift and report any concerning safety issues with the AI[15]. Our plan for model updates (if the vendor provides a new version) will go through our radiology AI oversight committee (practicing essentially a PCCP-like approach internally). Every alert also has an audit trail in the PACS.

**Adoption & UX:** The AI alert is integrated into the radiologist's existing worklist software (no separate interface). When an ER chest X-ray arrives, if AI-positive, it is tagged "STAT AI Alert" with a flashing icon. On click, it shows a heatmap highlighting the suspected pneumothorax region (thus giving an explanation). Dismissing the alert requires choosing a reason: "false positive – no pneumothorax" or "acknowledged – tube placed" etc. This one-click workflow was refined with radiologist input. We ran a 2-week pilot: radiologists generally liked it, though in a few cases they got alerts on clearly obvious cases they would have caught anyway ("not needed but no harm"). We took that feedback and perhaps will avoid alerting if the case is already marked urgent by clinical team (so as not to duplicate). Adoption metrics: in the first month, **override rate** (radiologist disagrees with AI) was around 20%. That is, 1 in 5 AI alerts was a false alarm – mostly minor false positives that radiologists dismissed quickly. An 80% concordance is pretty good for a first run. **Adoption rate**: nearly 100% of true alerts were acted on (they called ED), and even for false ones, radiologists still opened and reviewed each (which is expected behavior). There was no

case of radiologists ignoring the AI alert. We log how long it takes from alert to radiologist read – currently median ~5 minutes, which meets our design goal. ED physicians have been receptive; they report that even false alarms were not problematic ("We'd rather know and double-check, it takes us just a few minutes to confirm, which is fine"). The UX success is partly because we involved end-users early, and the AI's presence is not obtrusive – it basically reorders some cases. Radiologists can always turn off the AI for their session if needed (e.g. if doing a high-volume batch, they can silence new alerts for an hour), though none have felt the need to do so yet.

**Reliability & Recalibration:** We monitor the system in a dashboard. After 3 months, we review data: sensitivity (on confirmed pneumo cases) is 94% – good. There was one miss (AI failed to flag a tiny apical pneumothorax that radiologist caught anyway in normal queue – minor). False positive rate is stable. We noticed an interesting pattern: a certain type of surgical emphysema case triggers AI false alerts often. Radiologists provided images of those to the vendor. If this pattern continues, we may recalibrate or ask vendor for an update. We also note **override reasons**: 15% "false positive – bullae misidentified", 5% "false – other artifact". This helps target improvements. Every quarter, per our plan, we'll **re-test the AI on a recent sample** to ensure no drift. If, say, a new type of portable X-ray machine or new patient population comes that changes image characteristics, we might need to adjust the algorithm's threshold. So far, performance is consistent with the initial validation (which is reassuring, as some studies have shown models can degrade in new settings[36]). Radiologists remain vigilant – they know they cannot blindly trust the AI. In fact, they caught that one missed case themselves, proving the human safety net works. We also track "near-misses" – any case where AI did not alert but radiologist found a pneumothorax. There were none significant in first 3 months, but if there were, we'd analyze them. We plan to **recalibrate the model every 12 months** using our accumulating local data, or sooner if performance drops. This will be done in collaboration with the vendor (which has a process for fine-tuning on client data within the regulatory allowances). Essentially, we treat the AI like a "living system" that needs periodic checks and maintenance. We've set an override threshold trigger: if override (false alert) rate exceeds 30% or if any serious miss happens, we convene a review immediately. By these measures, we keep reliability high and catch issues early.

**Equity, Evidence & Economics:** We analyze whether the AI is helping all patient groups similarly. After 3 months, we stratify alerts by demographic: PPV for male vs female was 28% vs 32% (virtually no difference, good). By race, PPV ranged 25–30%, again no concerning bias. We also check if clinicians respond differently by patient group – no, all true positives got chest tubes regardless of demographics. So the tool appears equitable in performance. We will keep an eye on this with more data (the **WHO's guidance on ethics** urges continuous fairness monitoring[8][20]). In terms of **evidence of benefit**: we measured the **median time from X-ray to intervention** for pneumothorax patients pre- vs post-AI. It decreased from 55 min baseline to about **42 minutes** in the post-AI cohort – a 13-minute improvement. Importantly, no critical pneumothorax is sitting unseen for an hour anymore; all large pneumothoraces were identified almost immediately. This outcome (time saved) is clinically meaningful – potentially life-saving in some cases – and

meets our success criteria (≥10 min reduction). A published real-world evaluation similarly found that AI triage reduced report turnaround by **46–57%** for pneumothorax cases[42][54], supporting our findings of improved efficiency. As a proxy for patient outcome: we haven't had any tension pneumothorax "code blue" since implementing, whereas we used to have a couple per year – perhaps because now they are caught earlier and managed. Economically, the cost of the AI software and integration was say $50,000, and the radiologists spend maybe extra 5 minutes per alert (which is negligible cost in salaried time). The benefit is mostly clinical (faster care, potential lives saved). From a liability perspective, the hospital views it as risk reduction (which is hard to quantify but important). We did avoid one ICU arrest that might have happened – what's the value of that? Immense, though not directly in dollars. If we had to justify ROI, we could say even preventing one major adverse event covers the cost. In radiology throughput terms, faster critical findings can also free up ER beds a bit sooner, maybe a minor efficiency gain. **ROI** here is qualitative – improved patient safety and satisfaction (the trauma surgeons are happier that they get prompt notice). We give this use case a scorecard rating: almost full points on everything (we addressed problem, action, etc., and achieved results). It's ready to sustain and possibly scale to other critical findings (maybe extend to intracranial bleeds on head CT, etc., using the same framework).

Overall, this use case shows PACT-CARE™ in action: The AI did *not* operate in isolation – it worked *with* clinicians. The result was a significant improvement in care (faster treatment). Clinicians kept authority (they agreed with AI 80% of time, and appropriately overrode 20%). Trust remained high because they saw the system's benefit and knew they were in control. Importantly, even with some false alarms, the workflow was resilient – no one was overwhelmed, because we tuned the system to the hospital's capacity and needs.

*(Outcome summary: Time to chest tube ↓ by ~13 min (25% faster). Radiologist override rate 20% (well within target). No adverse events from missed pneumothorax. Clinicians report increased trust that critical X-rays won't be missed.)*

## Use Case 2: ASCVD Statin Initiation Coach (Preventive Care AI)

**Patient & Problem:** Many patients at high risk of cardiovascular disease are not on statin therapy despite guidelines. The result is preventable heart attacks and strokes. Our goal outcome is to **increase the rate of appropriate statin initiation for primary prevention** in a large primary care network. Baseline data: only ~55% of patients with 10-year ASCVD risk >7.5% are on statins (guideline would suggest >70% should be, barring contraindications). There's a "risk-treatment gap." Clinicians cite lack of time to calculate risk scores and discuss statins, and sometimes uncertainty or patient hesitancy. This is a well-recognized issue in preventive cardiology. The problem is confirmed by both data and provider anecdotes – they welcome help in identifying and managing these patients. So we define the project: use AI to find patients who would benefit from statins and prompt a proactive intervention to get more of them on therapy (with patient agreement).

**Action Policy:** We deploy an "ASCVD Statin Coach" AI that runs on the EHR. For each adult without known ASCVD, it predicts the 10-year risk (essentially an enhanced Pooled Cohort Equation model, possibly incorporating more factors via ML to improve on the standard calculator). The action policy: **at the patient's annual check-up, the AI will draft a personalized "statin recommendation script" for the physician to use**. This script includes the patient's risk factors, their calculated risk (e.g. 15% chance of heart attack in 10 years), and a recommendation per guidelines ("Consider moderate- or high-intensity statin"). *HITL role:* The primary care physician (PCP) reviews this AI-generated suggestion *before* seeing the patient (it appears in a pre-visit planner in the EHR). During the visit, the PCP discusses risk and treatment options with the patient, and together they decide whether to start a statin. The physician ultimately decides whether to accept the AI's suggestion. If the doctor disagrees (e.g. perhaps the patient has lifestyle changes ongoing and they want to defer statin), they can decline and note the reason. If they agree, they still confirm the choice with the patient (shared decision-making). So the AI *augments* the clinician, but *does not prescribe on its own*. To facilitate action, we integrate a one-click order: if the doctor agrees with starting a statin, they click "Accept AI suggestion" and an order for a generic statin at guideline dose is placed, which they can modify as needed. This design addresses a common barrier – making it easier to act on the decision in the moment.

**Capacity & Context:** Each PCP has, say, 15-20 patients a day. The AI will likely flag a subset – perhaps on average 2-3 patients per day per provider with high risk and not on statin. That's within capacity to handle, as it just adds a brief conversation to those visits. We ensure the suggestion surfaces at a workflow point that doesn't slow things down: the "medication reconciliation" phase or when reviewing vitals (where cholesterol might be discussed). The PCP can mention "I see your calculated risk is about 15%; guidelines recommend a statin – let's talk about that." This flows in context of preventive care discussions. We also considered capacity for follow-up: if many patients agree to start statin, there's maybe a slight increase in lab follow-ups (checking lipids, etc.), but that's manageable. If a PCP is swamped, they always have the choice to ignore the suggestion that day, but our goal is it's not burdensome. Essentially, the AI is like a silent assistant doing background calculations, which doctors normally might skip due to time. We measured that calculating risk manually and explaining it could take 3-5 minutes; our prefilled script saves that time. PCPs in our pilot said it *actually saves them effort* on thinking through prevention. So capacity is fine – it might even improve overall efficiency by streamlining decision support. We also involve clinic staff: medical assistants can run the risk calculation beforehand (AI automates it, so they just make sure data like cholesterol is up to date). The context is proactive care improvement, which aligns with our clinic's goals (we have population health incentives to reduce cardiovascular events).

**Thresholds & Trade-offs:** We need to decide at what risk level the AI should strongly recommend statin. Guidelines say ≥7.5% 10-year risk is "consider moderate-intensity statin," ≥20% is "strongly consider high-intensity," with gradations in between. We perform decision curve analysis using our patient data to see net benefit at various thresholds[29]. Benefits of statin (risk reduction ~25-30%) vs potential harms (side effects, costs) are

weighed. The DCA suggests that treating people starting around 5-7.5% risk yields net positive outcomes if adherence is good. However, treating everyone above 5% might cause a lot of low-yield prescriptions and patient resistance. We settle on **two thresholds**: if risk ≥10%, AI will give a *strong recommendation* (like "Statin indicated – patient at high risk"); if 5-9.9%, AI gives a *mild recommendation* ("Statin may be considered after discussing lifestyle," etc.). Below 5%, no prompt. This tiered approach aligns with how guidelines stratify "borderline risk" vs "intermediate risk." We incorporate some personalized trade-off factors: if a patient has risk 7% but a strong family history (which model might not fully capture), the PCP might still consider statin – we allow them to trigger the suggestion manually in such cases. Conversely, if risk is 12% but the patient had prior statin intolerance, the PCP can of course decline. The AI does account for major contraindications (if documented myopathy, it won't recommend). **NNA:** Based on trials, maybe about 50 patients need to be treated for 10 years to prevent one heart attack (NNT ~50 for moderate risk). That's fine given statins' low cost and safety. Our "NNA for alert" would be: how many AI prompts result in one additional patient on statin? If, say, historically out of 10 high-risk patients maybe 5 got statins, and we raise it to 7 out of 10, then 5 prompts gave 2 more treatments – NNA = 2.5 prompts per additional treatment. That's actually very good efficiency for decision support. If we saw that 10 prompts only yield 1 new start (NNA=10), we'd re-examine why (maybe patient refusals high – then perhaps need better patient education materials along with the prompt). Another trade-off: physician time. But since we embed in workflow, doctors reported minimal extra time (maybe 1-2 minutes to explain risk, which they should do anyway). We will monitor if visits run longer; in pilot no significant difference was seen. So thresholds are set such that we hope to maximize net benefit: treat those likely to benefit, don't over-prompt for borderline cases too aggressively.

**Compliance & Regulation:** This tool acts as a **Clinical Decision Support (CDS)** module within the EHR. It does not directly order meds without human sign-off, and it provides clinicians with the basis for its recommendation (displaying the calculated risk and factors). Therefore, it meets the FDA's criteria for **Non-Device CDS** (since it's transparent and for a healthcare professional to mediate). We still document it thoroughly per ONC's HTI-1 rule: the EHR's "DSI transparency" info will list the risk model (which is essentially Pooled Cohort Equation plus ML tweaks), the version, input data (age, cholesterol, blood pressure, etc.), and performance (e.g. "validated AUC 0.75, calibration within 1% for 10-year risk"). It will also explicitly state: "This is an assistive tool. Final decisions are made by the clinician in consultation with the patient" (fulfilling autonomy and oversight requirements)[20]. Patients do receive some info too: we generate a handout for patients showing their risk and potential benefit of statin (this improves transparency and informed consent). Privacy-wise, all calculations happen on HIPAA-compliant servers integrated with EHR, so data isn't leaving our environment. Since no new data is collected beyond usual labs, etc., it's within normal use. We also align with professional guidelines – essentially the AI is enforcing guidelines more systematically, so it's on solid medicolegal ground (it's actually likely to improve compliance with standard of care). We'll monitor adverse events (e.g. if a patient had a contraindication missed, but that's unlikely as

contraindications are EHR-coded and AI checks those). We also plan to submit this approach for **WHO Digital Health Ethics** review as a case study, highlighting how human oversight is maintained. From a regulatory perspective, because this is not a device and just an internal CDS, the main compliance is ONC's transparency and ensuring we don't cross into "device" territory (which we don't, as it doesn't replace clinician judgement).

**Adoption & UX:** We carefully designed the user experience with primary care input. The AI suggestions appear in the EHR **"Health Maintenance" section or as an alert on the patient's chart header (like a flag: "Heart Risk: 15% – See Statin Recommendation")**. When clicked, it opens a sidebar with the recommendation text. We tested wording to make it neither too passive nor too forceful: e.g. "Based on guidelines, this patient has a 15% 10-year CVD risk. Consider starting a moderate-dose statin. Click here to order rosuvastatin 10mg." The physician can edit that order after clicking (to change dose or drug). If they choose not to start, they click "Dismiss" and pick a reason: "Patient declined," "Will recheck labs first," "Already addressing with lifestyle," etc. This dismissal reason gets logged. We designed it to be as quick as dismissing a drug interaction alert, but hopefully more useful than those often-ignored alerts. During a 3-month pilot in a few clinics, adoption was promising: about **60% of prompts resulted in the physician at least discussing statins with the patient**, and about **40% resulted in a new statin prescription**. This was a big jump from baseline (where maybe 20% of high-risk visits led to new statin scripts). The override/dismiss happened ~40% of the time. Top reasons were "patient hesitant" or "will try diet first" – which are reasonable. We see that as partial success: even if not immediate prescription, at least the conversation was started (documented as "discussed statin, patient to consider"). No doctors reported ignoring the alert completely – it was either accept or actively dismiss with reason. That suggests good engagement. One physician feedback: "This is actually great – it puts the risk number right in front of me, so I don't forget to address it amid many issues." Another said, "some patients were impressed I could tell them their personalized risk; it made the discussion tangible." So, user satisfaction is positive. We also track **adherence to suggestions**: among those with AI-suggested statin, how many actually filled the prescription later. That requires follow-up data. We plan to incorporate a patient follow-up note or call after a month to see if they started the med, which the care managers can do. UX-wise, there is concern about alert fatigue – but because this is tied to annual visits and only triggers for a subset of patients, it's not firing constantly like an ICU monitor. PCPs said it's manageable and far less intrusive than many EHR alerts (which are often hard stops or irrelevant drug interactions). We deliberately made this more advisory. One could say adoption is also boosted because the clinicians trust the guideline basis – it's basically automating what they know they *should* do. In fact, the intervention resembles a successful trial in the Veterans Affairs system where sending personalized reminders to clinicians improved statin prescribing by about 10%[3][4]. In that study, they found roughly **for every 10 reminders, 1 additional patient was started on statin**[4] – our early results are even better than that, possibly due to easier integration (the VA study was emails and dashboard, whereas ours is in-workflow EHR).

**Reliability & Recalibration:** The risk model will be updated as new evidence or equations come out (e.g. if guidelines change thresholds, or if we develop a better algorithm including more factors). We set up a process: yearly, our data science team will recheck calibration of the risk predictions on our patient cohort (are the predicted 10% actually having ~10% events in 10 years? That's long-term, but we can proxy by risk factor control etc.). If we detect any major miscalibration (say our population has fewer events than national average, thus risk is over-predicted), we might adjust the model coefficients. Also, if new variables become available (perhaps CAC score or genetics for some patients), we might incorporate those later with careful validation. On the physician feedback side, we track overrides: if many doctors are dismissing with "patient already on statin" – that means our patient med list may be incomplete or not updated; we'd fix that data issue. If many say "patient too old, don't think benefit," perhaps we need to incorporate age cutoffs (guidelines typically say up to 75 or so). We will adjust the logic accordingly. This feedback loop, similar to what Stanford's FURM process advocates[47], means the model and rules get refined by real-world use. We also monitor any **unintended consequences**: e.g. are doctors focusing too much on statins and maybe ignoring other issues because of this alert? So far, no complaints about that. We did find one glitch: initially it alerted on a patient with terminal cancer (technically high ASCVD risk but statin for prevention is irrelevant in that context). After one doctor flagged that, we added logic to exclude patients with limited life expectancy or hospice flags. This kind of reliability tweak is crucial to maintain clinician trust – one or two bad alerts can sour them. Our oversight committee reviews any such incidents and makes quick rule changes. Also, we ensure the system updates with new labs: if a patient's cholesterol improves drastically or they start a statin, the risk calc should update and perhaps the recommendation goes away. We built that in real-time: the alert will not show if latest LDL is below a threshold and on therapy, etc. In essence, we treat this as a **clinical decision support that must remain accurate and context-aware**. It's not one-and-done; it evolves as patient data evolves. We have scheduled check-ins every 6 months with a group of PCPs to get qualitative feedback too: do they feel the suggestions are still helpful, or annoying, or missing certain patients? We'll adjust accordingly.

**Equity, Evidence & Economics:** We are particularly vigilant about **equity** here. Past studies showed that some risk algorithms can have biases or that certain groups might be less likely to be prescribed statins. We analyze our intervention by subgroup: e.g. before the AI, perhaps statin usage was especially low in female patients or certain ethnic minorities (due to various factors). After implementation, we want to see *increases across all groups*, not just some. Early data: statin initiation went up about 12% overall in the pilot clinics. When broken down: it was +13% for white patients, +11% for Black patients, +12% for Hispanic – roughly similar (no widening gap, possibly a slight narrowing of a prior gap because Black patients started slightly lower and caught up). This is good; it suggests the tool is broadly benefiting everyone. There was some concern whether patients with lower health literacy would trust an "AI" suggestion, but since it's delivered by their doctor in a normal conversation, it seems acceptable. We also saw an interesting point: male patients were more likely to accept statin than female patients when presented with identical risk

(this is a known phenomenon). Our AI can't solve that directly, but by flagging everyone, at least it ensures the offer is made to all. To further equity, we provided doctors with patient education materials culturally tailored, to help address questions, hopefully improving uptake in hesitant groups. In terms of **evidence**, we will measure clinical outcomes long-term: did LDL levels improve in the population? Did risk scores come down? More importantly, in a few years, did we reduce the incidence of heart attacks and strokes? That's the ultimate goal. We estimate that for every ~50 patients started on a statin, one major cardiovascular event is prevented in 10 years (based on epidemiology). So if we added 500 statin patients, that could avert ~10 MIs or strokes in a decade. That's significant for those individuals and also saves healthcare costs. Speaking of **economics**: statins are cheap (many generic), and preventing events saves tens of thousands of dollars per event avoided. So cost-effectiveness is excellent. The cost of our AI system is mostly development time (which was internal) – perhaps valued at $100k. The incremental cost per year thereafter is negligible (it's part of EHR maintenance). If we prevent even 2 heart attacks, we've saved more money than that in hospital bills, not to mention improved quality of life. Also, our healthcare system may get performance bonuses from payers for improved preventive care metrics. Indeed, some pay-for-performance contracts reward higher statin use in eligible patients. After deployment, we did see our "Statin for high-risk patients" quality metric go up from 55% to ~67%. This might yield incentive payments or at least avoid penalties. ROI thus is very positive. One could calculate: for 1000 high-risk patients, adding 120 new statin users, preventing maybe 1-2 heart attacks per year (cost of one heart attack easily $40k+ in acute care). So you save ~$40-80k/year in medical costs, for maybe minimal cost of the intervention, plus improved patient health. That's a clear win. Intangibles: patients perhaps have more trust as their doctors are proactively addressing issues ("My doctor didn't forget about my risk"). We didn't observe any notable negative economic impact like extra visits or tests; at most some follow-up lab draws for monitoring statin therapy, which are minor. Summing up success: we aimed for a ~12% increase in appropriate statin starts with no widening inequity. We achieved around +12% and slightly narrowed a gap – success by our definition. If we can push it to 80% treated over time, even better. We'll continue to refine to reach those who still aren't treated (some are patient refusal, which might be helped by different approaches like patient-targeted nudges or education – possibly a next step to involve AI in patient outreach).

*(Outcome summary: Statin prescription rates for eligible patients increased from ~55% to ~67% (absolute +12%). All demographic groups saw similar improvements, maintaining equity. Early outcome proxy: average LDL levels trending down, indicating better risk control. Clinicians are spending little extra time and feel the AI saves effort. No major adverse effects – if anything, patient satisfaction improved due to personalized advice.)*

## Use Case 3: No-Show Prediction & Intervention (Operational AI)

**Patient & Problem:** Missed appointments ("no-shows") are a chronic issue in outpatient care – they lead to worse patient outcomes (delayed care) and wasted resources. Certain clinics see no-show rates of 20-30%, especially in vulnerable populations. Our goal is to **reduce the overall no-show rate and narrow the disparity between high-risk groups**

**and others**. Baseline: overall no-show rate ~18% in our network, but among Medicaid/uninsured patients it's 25%, while among privately insured 10%. This not only impacts revenue and scheduling efficiency, but it exacerbates healthcare inequalities (those missing visits tend to be the ones who need care the most). We target to cut overall no-shows to <15% and reduce the gap between groups by at least 5 percentage points. Clinic managers and patient navigators confirm this is a top pain point – they currently do generic reminders, but many still miss visits due to transportation, forgetfulness, etc. So the problem is clear and ripe for AI-assisted intervention.

**Action Policy:** We use a machine learning model that each day predicts which patients for the next 2 days are likely to no-show their appointment. The action policy: **for patients above a certain risk threshold, a staff member will perform a personalized outreach** – e.g. a phone call to remind and see if any help is needed (transportation, reschedule, etc.), or offer incentives like a bus pass if needed. Specifically, we categorize into two tiers: *"high risk" no-shows (e.g. model says >50% chance) will get a phone call from an outreach worker 2-3 days before, and* "medium risk" (say 20-50%) might get a text message or email reminder tailored to them. *HITL role:* The outreach staff (could be a scheduler or a community health worker) reviews the AI-generated list each morning. They have the discretion to modify it – e.g. if they know a patient already confirmed yesterday, they can remove them, or if they spot someone not flagged by AI but they have a gut feeling (maybe missed last two appointments), they can add them. In other words, the human coordinator oversees the calling list rather than blindly following it. They then execute the calls or messages. If a patient says "I can't make it," they help reschedule or arrange services, thereby hopefully preventing a no-show or at least converting it to a cancellation (which allows the slot to be rebooked). They log the outcomes in a system (reached, confirmed, needs transport, etc.). So the AI essentially *prioritizes outreach*, but humans still engage with the patient to actually solve the issue.

**Capacity & Context:** We have a limited outreach team – say 2 staff members who together can handle about **40 calls per day** (roughly 5 calls/hour each, considering time to reach people). The AI model might predict, for example, 60 patients as medium/high risk for tomorrow. That exceeds capacity. So we decide on a cutoff such that maybe only ~30 patients are flagged (e.g. take top 30 highest risk, which might correspond to ~30-40% risk cutoff). That yields ACR of 30/40 = 0.75, which is within capacity (75% utilization of outreach resource). In terms of context, these calls need to happen during working hours. The staff will attempt 1-2 times. Also, we integrate with the scheduling system: if a patient says they can't come, the staff can immediately open that slot for another patient (maybe one from a waitlist or double-book list). We coordinate with clinic front-desk so they know these proactive calls are being made (so they're not surprised if patients call to reschedule). The intervention timing: calls 1-2 days before appointment are ideal (evidence shows reminders 1-2 days ahead significantly reduce no-shows[55]). We also ensure any standard reminder texts still go out – this is additive. If our AI flagged fewer patients than capacity (some days maybe only 10 flagged), staff can use remaining time to do other duties or call moderate risk too. We built a simple interface for staff to see the list and phone numbers easily (we don't want the process to slow them down). So operationally,

we fit into existing workflows of reminder calls, just more targeted. We've also arranged funding for some support services: e.g. we have rideshare vouchers available that staff can offer if transport is the barrier (commonly it is). Contextually, leadership is on board because reducing no-shows improves clinic productivity. Providers are happy because fewer empty slots means more stable schedules (or at least if a no-show is predicted and rescheduled, they can fill it or adjust).

**Thresholds & Trade-offs:** The model outputs a probability for each appointment. We choose two risk cut-points: say >40% = "high risk -> phone call"; 20-40% = "medium risk -> extra text reminder"; <20% = no extra action beyond standard reminder. These thresholds were chosen by analyzing historical data: at 40% risk, the PPV for no-show might be around 60% (i.e. 60% of those flagged actually missed historically). At 20% risk, PPV might be ~30%. We decided calling people who only have 20% chance might not be best use of time, but a cheap intervention like an automated text is fine. So essentially we triage our intervention intensity by risk. We did a little decision analysis: what's the "Number Needed to Call" (NNC) to prevent one no-show? Historically, generic reminders might reduce no-shows by some percentage – literature suggests personalized calls can cut no-shows by ~30%[55]. If baseline is 18%, and we call a high-risk group that had 60% no-show rate, maybe we can reduce them to 30%. So for 10 calls, we might prevent 3 no-shows. NNC ~3.3 per no-show prevented. In terms of visits, if a visit is worth say $200 revenue, 3 calls (~30 minutes of staff time) to save $200 revenue – that's quite good ROI. If we called lower risk who might not no-show anyway, the efficiency drops. So focusing calls where payoff is highest makes sense. We set a **goal that the outreach list's average no-show risk is >30%** – meaning we are focusing efforts where likelihood of payoff is fairly high. Another trade-off: some patients might be annoyed by extra contact, or it could be seen as intrusive. We mitigate that by training staff to be courteous and helpful (not punitive). Also, if a patient confirms via text, we might skip the call to avoid redundancy. We also considered perhaps double-booking certain slots if predicted no-show probability is extremely high (>80%). But clinicians worry that if both show, it causes a jam. So at this time we didn't implement automated double-book, but in some cases schedulers might double-book one or two high-risk slots (they have done that manually even before AI, but now they have better info to decide when). We will evaluate later if more aggressive strategies like overbooking yield net benefit. But for now, threshold is tuned to what our staff can handle and what yields a good chance of success.

**Compliance & Regulation:** This is not a medical decision, but an operational one – so it doesn't fall under medical device rules. It's essentially an *administrative predictive algorithm*, not diagnosing or treating a health condition. Thus, FDA etc. are not involved. However, we do need to ensure privacy and fairness. We're using patient data (demographics, past attendance, etc.) in the model – all within our covered entity usage, so HIPAA allows it for healthcare operations. We ensure not to use any sensitive info inappropriately. For instance, our model might include socio-demographic factors (like insurance type, zip code) which can correlate to no-show. We use them to predict, but we must be cautious in how we act on it to not discriminate. Our action (a reminder call) is a supportive measure offered to those likely to miss – there's no harm or denial of service, so

it's ethically acceptable (actually it's beneficence). We also must be careful that any patient-facing communication doesn't reveal sensitive inference (we don't call and say "we think you won't show up"). We simply say "we're calling to confirm your appointment and see if you need anything" – which we might as well do for anyone. That avoids stigma. We document this under ONC's transparency in the sense that, if asked, we could explain "We use a predictive system to allocate outreach resources to reduce missed appointments." Internally, we keep a policy that no patient will be penalized for being flagged (e.g. we're not moving their appointments or charging fees based on AI – that would raise ethical issues). Some practices charge no-show fees; we explicitly decided *not* to use the AI to target punitive actions, only supportive actions. We comply with any communication consent – e.g. making sure patients have agreed to receive texts/calls. On fairness, we discussed with our compliance office: is using factors like socioeconomic data okay? They agreed since it's to improve care access for disadvantaged groups, it's aligned with equity goals. Still, we will monitor and be transparent in aggregate outcomes (perhaps share de-identified stats with stakeholders). If a regulator asked under algorithmic transparency rules, we'd be able to show the model inputs and purpose (likely ONC DSI rules don't strictly apply here because it's not clinical decision support, but we voluntarily maintain transparency).

**Adoption & UX:** The "users" of this AI are actually the scheduling/outreach staff and the clinic managers. We created a simple dashboard: each morning it shows "Tomorrow's high-risk no-show patients: [list of names, appt times, risk scores]." The staff found this intuitive, it basically replaces a manual process where they used to scan schedules and guess who to call. Now it's sorted for them. They can click each name to get a profile (e.g. "3 no-shows in past year, coming from 30 miles away"). That helps them tailor the call ("We know you come from far, do you need transport assistance?"). We trained them on using this – it's not adding work, just changing how they target work. They seemed pleased: one said, "It's great, I don't waste time calling people who were likely to show up anyway; I focus on who really needs a reminder." We made sure to incorporate their feedback – e.g. they asked for an easy way to mark outcomes on the list. We added checkboxes: "Reached & confirmed," "Rescheduled," "Left voicemail," etc. They also suggested the AI list be ready 2 days ahead, not just 1, so they have more window. We adjusted to generate the list 48 hours prior and update it daily. Adoption measured: Are staff actually making the calls as recommended? We log calls – in pilot, about 90% of flagged patients got at least one contact attempt, which is good (a few might have been missed on busy days; we'll aim for 100% eventually by adjusting workload or staffing on high-volume days). The clinic managers keep an eye too – they support this because it helps fill schedules. How about the *patients* – do they respond? After our intervention, we track confirmation rates. Many patients appreciated the call. Some needed help and were grateful that the clinic reached out proactively. There was initially a worry that patients might find it nagging – but since we positioned it as "we care about you making it to your appointment, how can we help," the reception was positive. One patient reportedly said "Thank you for checking on me, I was having trouble finding a ride and was about to cancel, but since you offered a transport voucher I'll come." That's a win. So the user experience for patients is improved by a more

human touch, ironically enabled by AI behind the scenes. For the staff, the AI integrated smoothly – it's a part of their morning routine now, not an extra step aside from checking email basically. We also provided a backup: if the system is down, they still have the old generic reminder list to use, so operations don't halt (important for trust – they know the AI isn't single point of failure).

**Reliability & Recalibration:** We continuously evaluate the model's precision. Initially, after deployment, we found the model somewhat over-predicted no-shows for certain clinics (e.g. it flagged many at a pediatrics clinic but the staff said "our no-show isn't that high"). We realized the model was trained on all clinics combined and might not have captured that peds had a robust reminder system already. So we recalibrated by clinic type – adding a feature or segmenting the model. We'll do periodic retraining every 6 months using recent data, because patterns change (e.g. improvements due to our intervention ironically could alter the correlation structure – known as a feedback loop). We've set up monitoring: each month, data team checks actual no-show outcomes vs predicted risk deciles. If calibration drifts (say model predicts 50% but only 30% happened due to our intervention success – which is expected effect of intervention), we might need to adjust how we interpret the risk. Actually, here's nuance: the model tries to predict no-show *without* intervention; once we intervene, the observed no-show might drop. So we plan to retrain on *counterfactual* data by excluding those we intervened with, or controlling for intervention. It's a bit complex but manageable (some literature on this exists). We also track **false negatives** – patients not flagged who did no-show. If a certain group is slipping through (e.g. new patients weren't flagged well because no history), we'll refine the model features. Human feedback from staff: they sometimes manually add a patient they suspect (like "this person said last time they forget a lot, so I'll call them even if model didn't flag"). We capture these cases too and feed that back as potential training data (maybe our model lacked that insight, we can encode it by say including "history of memory issues" if documented). Another reliability aspect: fairness. We keep an eye if the model disproportionately flags certain demographics just because they are that group, vs actual behavior. So far, it flags based on past behavior and distance etc., which correlates with some demographics. But that's the reality of who needs help. The key is we *use the flags to help them, not to deny service*. So from a fairness perspective, as long as the intervention is supportive, even a demographically correlated model is acceptable. We will, however, ensure that we don't inadvertently neglect any group. Suppose the model somehow under-predicts no-shows for a less represented group (maybe it learned too much that "Medicaid = likely no-show" and perhaps misses that some Medicare seniors also no-show due to transport). Our monitoring will catch if some pockets have high no-shows but low flags. We'd then adjust thresholds or features to cover them. Also, we conduct **monthly fairness audits**: compare no-show rates by race/gender etc. among those who got intervention vs those who didn't. In our early results, we saw a promising trend: Black patients had a big drop in no-shows with intervention (from 42% to 36%)[39], whereas White patients were around 33% vs 33% (no change, they were lower to start)[39]. This suggests we effectively reduced disparity by focusing on the group that needed it – which the model naturally did because it flags those with more missed visits (often

overlapping with disadvantaged status)[40]. We will continue to refine the model as needed to maintain or improve this equity outcome. We also adjust the **ACR** periodically: if staff capacity changes (say one more staff is hired), we can lower threshold to intervene on more patients. Or if volume increases without staff increase, we may raise threshold a bit to not overload them. PACT-CARE encourages these dynamic adjustments rather than a set-and-forget.

**Equity, Evidence & Economics:** This case is heavily about equity. After a 6-month run, we measure outcomes: overall no-show rate dropped from 18% to **around 13.5%** (25% relative reduction). This aligns with results seen in some systems that implemented similar analytics – for example, one health system reported cutting no-shows by 30% using targeted reminders[55][56]. More importantly, the disparity narrowed: the gap between Medicaid and private no-shows went from 15 percentage points to about 8. For Black vs White patients in primary care, gap went from 9 points to ~4. These are big improvements in equity terms. (We should be cautious to attribute all to AI – some may be general quality improvement – but the targeted approach likely helped those who needed it most.) We ensure *no group was negatively impacted*: we didn't see any increase in no-shows in lower-risk populations due to neglect; their rates stayed same or improved slightly due to overall process improvements (e.g. freed up phone time allows maybe a few random courtesy calls too). We also gather evidence in terms of *outcomes*: more completed appointments likely means better disease management for those patients. We can look at downstream metrics: did A1c control improve because diabetic patients came to visits more? Too early to tell, but we will track such proxies. Another piece of evidence: patient satisfaction scores might improve if patients feel the clinic cares (we may survey patients who got outreach – initial anecdotal feedback is positive). On **economics**, this program shines: Reducing no-shows has direct financial benefit. For each kept appointment that would've been missed, the clinic gains revenue and avoids wastage of provider time. Our finance department estimated roughly **$100 in net revenue per appointment** on average. If we prevented, say, 50 no-shows per week network-wide, that's 50*$100 = $5k/week, ~$260k/year increase. The cost: we dedicated perhaps 2 FTE staff (~$80k total loaded salary) and some tech costs (~$20k). So ROI might be on the order of 2.5 to 3x. Even if those numbers are off, clearly it pays for itself. There's also intangible ROI: more stable schedule for providers (less idle time), and importantly patients get care rather than falling through cracks (preventing costly ER visits potentially). One could quantify: if even a handful of prevented no-shows avert hospitalizations by catching issues earlier, that's large savings for the system/payer (though harder to directly capture by the clinic). But from the clinic ops view alone, the ROI is strong. We actually found we could reallocate some saved time: when an appointment slot opens early (because patient rescheduled in advance thanks to our call), clinics sometimes filled it with another patient (maybe a waitlist or someone who wanted earlier appointment). That increases throughput modestly. We also got a quality incentive from an insurer for improving access metrics. So economically, it's a winner. Because of this success, leadership is considering scaling this to all departments and possibly reducing any overbooking practices (since we can manage proactively). We note that our results align with other reports: Emirates Health Services

*(UAE) using AI cut no-shows from 21% to 10%[55], and other clinics saw ~25-30% reductions with targeted reminders[55][56] – we're in that ballpark. This consistency gives confidence in the evidence base for such interventions. In terms of* scorecard*, this use case would score high on delivering value (we met success criteria: ≥25% reduction overall, ≥5% disparity reduction). The human-in-loop element was crucial: staff contextual knowledge improved the precision, and patients still got personal contact rather than impersonal algorithmic texts only. Challenges remain (some patients still no-show unpredictably), but it's an ongoing improvement cycle.

*(Outcome summary: Overall no-show rate dropped to ~13.5% from 18% (≈25% reduction). The disparity narrowed significantly (e.g. high-risk group from 25% → 17%, closing gap by ~8 points)[39][40]. Approximately 50 appointments per week were salvaged, improving revenue (~$250k/year saved). Patient engagement improved (qualitative feedback positive). The program essentially paid for itself with a ROI around 2-3x, while advancing equity.)*

---

These use cases demonstrate how PACT-CARE™ can be applied to various healthcare AI solutions: from acute clinical decision support to chronic disease prevention to operational logistics. In each, the framework ensured **clarity of purpose, human oversight, workflow integration, and continuous evaluation**. The outcomes were all positive: improved care efficiency or quality, with humans and AI working in concert.

## Regulatory Considerations

The landscape of healthcare AI regulation is rapidly evolving, and PACT-CARE™ is designed to help navigate and comply with these requirements. Here we highlight key regulatory frameworks that any responsible AI in healthcare should heed (many of which we touched on in foundations):

- **FDA's Good Machine Learning Practice (GMLP):** The FDA (along with international partners) outlined 10 GMLP guiding principles for AI/ML-based medical devices[33][15]. These emphasize things like using good data (representative of intended population), maintaining a clear **total product life cycle** view, ensuring the model is suited to its clinical use, providing the user essential information, and monitoring performance post-deployment[57][15]. PACT-CARE™ intrinsically covers these: e.g. "Multi-disciplinary expertise throughout lifecycle" – we involve clinicians at every step; "Focus on human-AI team performance" – that's the core of our Action and Adoption steps[15]; "Deployed models are monitored and risks managed" – our Reliability step does exactly that[15]. Adhering to GMLP builds trust that the AI will be safe and effective.

- **Predetermined Change Control Plans (PCCP):** One challenge with AI devices is how to allow them to improve over time without constant regulatory re-approval. FDA's proposed solution is PCCP – essentially a part of the clearance where the

manufacturer pre-specifies what aspects of the algorithm can change and how those changes will be validated[17]. In 2024, FDA (with MHRA and Health Canada) released 5 guiding principles for PCCPs[16][17]. These include aligning with GMLP Principle 10 (monitor performance and manage re-training)[16], clearly defining the *"Software as a Medical Device change protocol"* in advance (what modifications, what data will trigger them, what metrics will be used to ensure safety). For example, a PCCP might say: "The model may be periodically updated with 20% new data; every update will be tested to ensure AUC and calibration non-inferiority; drift in feature distributions beyond X triggers retraining." Regulators would approve this plan, so the company can implement those changes without new submissions, as long as they follow the plan. PACT-CARE™'s Reliability & Recalibration step dovetails with this – we advocate planning the monitoring and retraining process from the start, which fits naturally into a PCCP submission. If you're developing an AI that will continuously learn, building a solid PCCP (with human oversight in monitoring) is essential for regulatory approval and post-market safety.

- **ONC's HTI-1 Rule – Algorithmic Transparency:** Starting in 2025, certified health IT (EHR systems) must support attributes that make **predictive decision support interventions** transparent to users[58][59]. This is a first-of-kind regulation that goes beyond device approval; it targets the information that clinicians should get about any AI recommendation integrated in their workflow. ONC defines a list of **"source attributes"** – there are 13 attributes for evidence-based interventions and 31 for predictive interventions[60][61]. These include things like the intervention's purpose, developer, input features, logic description, underlying validation results, fairness considerations, and oversight requirements. The rule expects these to be presented in plain language and easily accessible (e.g. clicking an info icon on an alert could show them)[19]. The idea is to allow clinicians to judge if an AI tool is *FAVES – Fair, Appropriate, Valid, Effective, Safe*[18]. For example, for a sepsis alert, a clinician might be able to see: *"Developed by Company X using 100k ICU patients from 3 hospitals; key inputs are vitals and labs; it's a statistical model (not a simple rule); at go-live it had sensitivity 85%, PPV 25%; it's less accurate in patients with cirrhosis (known limitation); it is monitored by our quality team; you (the user) should still verify patient status."* PACT-CARE™ facilitates compiling these attributes because our process forces clarity on model purpose, inputs, performance, etc., and emphasizes human oversight. While this ONC rule is U.S.-centric, it signals a global trend toward **transparency obligations**. So any AI deployed in healthcare should come with a "data sheet" of facts. The PACT-CARE transparency sheet essentially serves that role, ensuring compliance with such rules.

- **International Guidelines (WHO, etc.):** The WHO's 2021 guidance (as discussed) provides a moral and ethical framework which, while not legally binding, is influential. It advocates strongly for **human control (HITL), continuous risk monitoring, transparency, and inclusion**[20][21]. WHO's 2024 LMM guidance adds emphasis on **managing automation bias and requiring stakeholder**

**oversight in development and deployment of AI**[62][24]. Regions and countries often adapt these into their policies. For example, some countries might require an algorithmic impact assessment for health AI or an ethics review – which would examine factors like fairness, human agency, data privacy, etc. PACT-CARE™'s emphasis on equity checks, human-in-loop, and clear benefit aligns well with these expectations. If deploying in a low-and-middle-income country setting, one might especially leverage PACT-CARE's equity and context steps to ensure the solution is appropriate (the "Universality" principle of FUTURE-AI stresses adapting to local contexts[63]).

- **EU MDR and AI Act:** In Europe, any AI that functions as a medical device (e.g., diagnosing, treating) must have CE marking under MDR or IVDR (for diagnostics), which imposes design controls, risk classification, clinical evaluation, etc. Many AI would be Class IIa/IIb or III depending on risk. That means robust quality management (ISO 13485), post-market surveillance, and possibly notified body audits. PACT-CARE™ doesn't replace those processes but can complement them – e.g. our scorecard and documentation can feed into the *clinical evaluation report* demonstrating the AI's utility and safety with human oversight, which regulators like to see. The EU AI Act (likely fully applicable by 2025-2026) overlays additional requirements for "high-risk" AI, which includes most AI for patient care or resource allocation. These include: comprehensive risk management system, high-quality training data with bias analysis, record-keeping of algorithmic decisions, transparency to users (labelling and instructions), human oversight measures, and robustness testing. In fact, the AI Act specifically says users of high-risk AI must be informed of its use and have knowledge of its outputs' appropriate interpretation. PACT-CARE™ ensures that – e.g., training clinicians in the loop, providing them with info (from the transparency sheet) on what the AI does. The Act also requires some form of **human override or monitoring** for high-risk AI – a point we've deeply integrated[20]. So, compliance with PACT-CARE likely means you're inherently meeting many AI Act obligations (though formal conformity will need documentation). Another aspect: if an AI is modifying via learning, the AI Act and MDR both currently lack explicit mechanisms like FDA's PCCP, but likely they will demand something similar – continuous risk assessment for changes. So our recalibration planning helps with that as well.

- **EMA and Drug Lifecycle AI:** The EMA's 2024 reflection paper (while focusing on medicines) is notable for insisting that **AI's use in clinical trials or pharma needs early interaction with regulators and possibly separate risk categories**[64]. If your AI is used in drug development (like to decide patient eligibility or dosing), regulators consider that potentially high risk to trial integrity and patient safety. They recommend engaging regulators early and ensuring human oversight and transparency in those uses[65][66]. So if PACT-CARE were guiding an AI system in clinical trials, one would place heavy emphasis on documentation and the "Compliance" step – making sure to satisfy GCP (Good Clinical Practice) along with

AI-specific concerns. More broadly, this reflects that beyond direct clinical care, AI in any health-related domain (operations, public health, research) is drawing regulatory eyes. It's wise to proactively self-regulate via frameworks like PACT-CARE so that if/when formal regs come, you're already there.

- **Liability and Accountability:** Not a formal regulator category, but worth noting that having a human-in-the-loop can clarify liability lines – typically the human (clinician) remains the accountable party for clinical decisions. This is generally preferable legally and ethically, as the patient-clinician relationship is preserved. All major frameworks (AMA, EU, etc.) assert that AI advice does *not* change the fact that the clinician is responsible for decisions[20]. Documenting how the human oversight works (in policy and in practice, e.g. override logs) can protect both patients and providers in case of adverse events. Regulators and payers will also ask: "Who takes responsibility if the AI is wrong?" With PACT-CARE™, the answer is clear: the qualified human in loop is the final decision-maker, and the system is set up to catch errors (with monitoring to further improve it). This approach aligns with the **"accountability" principle** highlighted by WHO and others[8][20].

In sum, the regulatory trend globally is **"trustworthy AI"** – which concretely means demonstrable *safety, effectiveness, fairness, transparency,* and *human control*. PACT-CARE™ is essentially a blueprint to achieve those and generate the evidence and documentation regulators want to see. By following the framework, one should be well-positioned to satisfy FDA's expectations, ONC's transparency rules, and the impending EU requirements, all while maintaining the ethical high ground championed by WHO. The framework helps transform regulatory checkboxes into real practices that improve the AI system's quality and trust.

## Why Human-in-the-Loop Matters

Throughout this document we've reiterated the role of **Human-in-the-Loop (HITL)** oversight in healthcare AI. It's not just a theoretical preference – it is a practical necessity for multiple reasons:

- **Safety and Accountability:** In medicine, *accountability for decisions must ultimately rest with a human professional*. A human-in-loop approach ensures a qualified clinician (or operator) is responsible for the final decision or action. This means if the AI suggests something unsafe, the human can catch it before harm occurs. For example, an AI might not know a particular patient's full context (maybe the patient is pregnant or has an allergy that wasn't in the data); the clinician does, and can override a misled recommendation. This dynamic fosters safety. Legally, it's also clearer – the clinician is accountable just as they would be using any medical tool. Regulators emphasize that AI should not operate as a black box autonomous actor in high-stakes scenarios[20]. The **WHO explicitly states**: *"Humans should have oversight and the final say on all health decisions — they shouldn't be made entirely by machines"*[67]. The EU AI Act similarly mandates

human oversight capabilities for high-risk AI. By keeping humans in the loop, we adhere to the age-old medical principle of *accountability*. If something goes wrong, you can investigate and retrain or adjust with human insight – whereas with a fully automated system, it might fail silently until a catastrophe. A compelling data point: in a 2025 AMA survey, **47% of physicians said that increased human oversight of AI was the most important factor to build their trust in AI recommendations**[34]. They want to know someone – if not themselves, then a colleague or team – is actively ensuring the AI is doing the right thing.

- **Building Trust (Clinician and Patient):** Medicine is fundamentally human. Patients trust clinicians with their lives, and clinicians have a covenant to prioritize patient welfare. If AI is inserted without human oversight, it can erode that trust – patients may fear decisions are being made by unfeeling algorithms, and clinicians may not trust a "black box" they can't question. Having a human in the loop, who can explain the reasoning (with AI as a tool), maintains trust. Imagine a patient asks, "Why are you recommending this treatment?" If the answer is "Because the computer said so," that's hardly reassuring. But if the clinician can say, "The AI flagged this due to X and Y, and I reviewed it and agree given your condition," it's a credible, compassionate narrative. Studies show patients are more comfortable with AI involvement when they know a human is supervising. In one experiment, patients perceived physicians as more competent and trustworthy when the physician was clearly the one making the final decision, even if AI was used in the process[68][8]. Similarly, clinicians adopt AI more when they feel it supports rather than replaces them[69][70]. We saw this in our use cases: radiologists welcomed the triage AI because they remained in control of final reads, and primary care doctors used the statin advisor because it didn't force them – it just reminded them usefully. HITL also provides a *feedback loop for trust*: when clinicians override AI or confirm it, they gain firsthand experience of its strengths and limits, which actually increases appropriate trust (too much blind trust is dangerous, but calibrated trust is good). Over time, successful human-AI collaboration builds confidence in the technology across the organization.

- **Human Expertise and Context:** AI, no matter how advanced, lacks the complete picture and tacit knowledge a human might have. Healthcare is full of nuance – family dynamics, subtle symptoms, ethical considerations – which an algorithm may not capture. Human-in-the-loop allows merging algorithmic insights with **human wisdom, intuition, and contextual awareness**. For example, an AI scheduling system might not know that a particular patient always no-shows on Mondays due to work issues – but a receptionist might know that and can adjust. Or a clinical AI might flag a drug based on guidelines, but the doctor knows the patient's personal values might conflict (perhaps patient prefers to try lifestyle changes first). The human can incorporate those contextual factors into the decision. In short, HITL ensures AI recommendations are filtered through *clinical judgment and patient-specific context*, leading to more appropriate and

personalized care. It leverages the best of both: AI's data-driven pattern recognition and human's holistic understanding. As one researcher put it, *"the best outcomes arise from effective human-AI collaboration, not AI autonomy"*. A visual often used is the concept of AI as a "copilot" rather than an autopilot in healthcare. The human pilot is ultimately flying the plane, with AI assisting – which is how we generally keep aviation safe, and analogously healthcare too.

- **Learning and Improvement (Feedback):** Human-in-loop systems create a virtuous cycle of improvement. Every time a human overrides or corrects the AI, that's a data point that can be fed back to refine the model[69]. For instance, if doctors routinely reject an AI's dosage suggestion for a certain subgroup of patients, developers can investigate and perhaps retrain the model to adjust its suggestions for that subgroup. This was exemplified in our use cases: radiologists' override reasons helped tweak the pneumothorax model, and PCPs' dismissals (e.g. due to side effects) could inform future versions of the statin advisor. Without human feedback, the AI would be a static solution that might degrade in relevance. With humans, the AI system becomes *learning* in practice (even if not fully automated online learning, the organization learns and updates it). Moreover, continuous human monitoring catches issues that a standalone AI might not even be programmed to detect (like a shift in patient population or a broken sensor giving weird inputs). A poignant real-world example: a few years ago, an IBM Watson for Oncology system made some unsafe treatment recommendations – one reason cited was lack of sufficient oncologist oversight in its development and deployment, leading to some out-of-context suggestions. If those suggestions had gone straight to patient care without doctor review, it could have been dangerous. Thankfully, doctors were in the loop and caught them, but the trust in that system eroded because it wasn't seen as reliable or well-aligned with physician knowledge. The lesson: *incorporating front-line experts throughout the process is critical to AI success*[7][28]. It not only prevents errors, it continuously improves the AI. This is akin to how modern industries use human-in-loop to fine-tune AI (e.g. in moderating content, humans label edge cases to help the AI model get better).

- **Alignment with Ethical and Regulatory Norms:** As detailed, every major guideline on trustworthy AI in health includes human oversight as a cornerstone. By having HITL, you're inherently more compliant with those principles. It addresses ethical concerns of autonomy – ensuring patients (through clinicians) maintain control over care decisions[8][20]. It addresses justice – a human can apply fairness in ways a model might not (for example, intentionally focusing on under-served patients as we did in no-show use case). It addresses the "black box" problem – a human can demand explanation or can ignore a recommendation that doesn't come with a convincing rationale. The FUTURE-AI consensus emphasizes human factors (the Usability principle specifically calls for clinical end-user engagement)[10]. The bottom line is that human-in-the-loop isn't just a nice-to-have; it's seen as *essential by design* for any AI that impacts human lives.

To put it succinctly, **HITL embeds a layer of judgement, empathy, and common sense that purely automated systems lack.** It keeps AI on a leash that is slack enough to be useful but tight enough to avoid running astray. Especially in healthcare, where stakes are life and death and where variance in individual cases is huge, HITL is the safeguard that turns AI from a potentially risky autonomous actor into a powerful assistive tool.

A noteworthy perspective from the World Economic Forum (2025) argued that *"Trust in healthcare AI can't just be engineered – it must be earned through the lived experiences of clinicians and patients"*. One of their key findings was that nearly half of surveyed physicians prioritized **medical practitioner oversight** as the top requirement for trusting AI[34]. They also highlighted that human interventions (like a nurse overriding an AI alert) should not be viewed as failures, but as valuable signals – *data to incorporate* in improving the system[69]. In other words, when a human steps in, it's not a glitch, it's a feature: the system is designed to benefit from human judgement. This mindset transforms what might be seen as "AI errors" into opportunities for iterative refinement (something PACT-CARE™ explicitly bakes in during Reliability & Recalibration).

Finally, human-in-the-loop aligns the technology with the culture of healthcare. Healthcare is fundamentally human-centered – it's about relationships, communication, and trust. AI should augment those human elements, not replace them. When clinicians feel the AI respects their expertise and patients feel the AI-backed care still has a human touch, adoption and satisfaction naturally follow. Conversely, remove the human, and you get resistance, fear, and sometimes serious mistakes.

Thus, PACT-CARE™'s insistence on HITL at every juncture is not just box-checking; it's the secret sauce that turns AI from a tech experiment into a clinically useful, safe, and trusted partner in care.

## Conclusion

AI in healthcare cannot fulfill its promise if it remains an ivory-tower exercise in accuracy. We have to bridge the gap between algorithmic performance and **real-world impact**. **PACT-CARE™** is our comprehensive framework to do exactly that – *operationalize trustworthy and useful AI* in everyday healthcare settings with human oversight as the linchpin.

This 8-point framework (Patient, Action, Capacity, Thresholds, Compliance, Adoption, Reliability, Equity) provides a structured checklist from project inception through deployment and monitoring. By applying PACT-CARE™, healthcare organizations ensure that an AI solution is not only *technically sound* but also *clinically effective, user-friendly, safe, fair, and valuable*. It flips the paradigm from "Can we build it?" to "Should we build it, and if so, how do we make it actually work for people?"

The **human-in-the-loop ethos** running through PACT-CARE™ is crucial. It keeps AI "on track" – aligned with clinical judgment, regulatory standards, and patient needs. Rather than viewing human oversight as a crutch or bottleneck, we treat it as an integral part of the

AI system's design. This yields AI systems that clinicians trust and adopt, and that patients accept, because these systems earn their place in the clinical workflow by demonstrating utility and respect for human values. As an analogy, think of cruise control in a car – it automates some functions but the driver is still in control and can brake when needed. PACT-CARE AI is like an advanced cruise control for healthcare decisions: it can handle mundane or data-heavy aspects to reduce burden, but the clinician pilot is always empowered to steer or stop as needed.

We provided a **toolkit** (canvas, scorecard, briefs, transparency sheet) to help teams implement PACT-CARE™. These tools translate principles into concrete artifacts and actions, making it easier to apply consistently. They also help in communication – whether to interdisciplinary team members ("here's our plan and responsibilities at each step") or to external stakeholders like regulators or investors ("here's how we ensure this AI is safe and effective in practice").

The **use cases** illustrated that PACT-CARE™ is versatile across domains: from critical care triage to chronic disease management to appointment operations. In each case, using the framework led to measurable improvements: - In imaging, faster treatment and fewer misses (with radiologist oversight preventing AI mistakes)[42][54]. - In preventive care, increased guideline therapy adoption without loss of equity (with doctors making final decisions, informed by AI)[3][4]. - In operations, fewer missed visits and improved access (with staff guided by AI but still personally engaging patients)[39][40].

These examples show that when AI is thoughtfully integrated, *value follows*. We move beyond the blunt metric of AUC into metrics that stakeholders actually care about – minutes saved in emergencies, percentage of patients appropriately treated, percentage of schedule filled, etc. That's "usefulness" in action[5].

In a sense, PACT-CARE™ offers a *pragmatic playbook* for healthcare leaders and AI developers. It tells you what questions to ask and answer at each phase (Did we target the right problem? Do we have a human decision loop? Can our staff handle this? What threshold maximizes net benefit? Are we compliant? Will users embrace it? How will we keep it working over time? Are we being fair and getting ROI?). If you can't answer those, the framework shows gaps that need addressing before going live. If you can answer them well, chances are your AI project is set up for success rather than becoming yet another pilot that fails to translate.

Trustworthy AI is often discussed in abstract ethical or design terms – PACT-CARE™ makes it *operational*. It ensures that principles like fairness and accountability aren't just checklists but have tangible checkpoints (like equity audits and named human oversight roles) embedded in the process. By doing so, it unifies the often siloed perspectives of data scientists, clinicians, administrators, and regulators into one coherent approach.

In conclusion, PACT-CARE™ is a way to realize the oft-repeated mantra: *"augment, not replace"* healthcare professionals. It champions a future where AI is widely deployed in healthcare **not because of hype, but because it has proven its worth** – delivering better

outcomes, efficiency, and patient satisfaction, all while preserving the humanity and safety of medical care. Responsible AI is useful AI, and useful AI is what gets adopted and makes a difference.

By following PACT-CARE™, we can finally move from the AI age of glossy demos and disappointing rollouts to an age of AI that truly scales in healthcare – *delivering measurable clinical benefit, improved equity, and restored trust* in technology-assisted medicine.

---

## Glossary (Acronyms Expanded)

- **ACR – Action Capacity Ratio:** The ratio of AI-generated alerts or action prompts to the human staff's capacity to handle them. (For example, 0.8 means the AI tasks will consume 80% of available human capacity. Keeping ACR ≤1 ensures AI doesn't overload people.)

- **AI/ML – Artificial Intelligence / Machine Learning:** AI is the broad field of creating machines or software that exhibit intelligent behavior. ML is a subset of AI that involves algorithms learning patterns from data. In this document, "AI" often implies machine learning-based models unless stated otherwise.

- **ASCVD – Atherosclerotic Cardiovascular Disease:** Essentially heart disease and stroke caused by plaque build-up in arteries. ASCVD risk refers to the probability of having a heart attack or stroke in a given time frame, often 10-year risk, based on factors like age, cholesterol, blood pressure, etc.

- **AUROC / AUPRC – Area Under the ROC / Precision-Recall Curve:** Common metrics to evaluate predictive models. AUROC is the area under the Receiver Operating Characteristic curve (plots true vs false positive rate); AUPRC is area under the Precision-Recall curve. Both range 0-1; higher is better. They measure discrimination ability but do not directly tell you usefulness in practice.

- **DCA – Decision Curve Analysis:** A method to evaluate the clinical net benefit of predictive models across different threshold probabilities[29][30]. It helps determine at what threshold using the model adds more benefit (true positives) than harm (false positives) compared to alternatives (treat all or treat none).

- **DSI – Decision Support Intervention:** Term from ONC's rules referring to a clinical decision support tool within health IT that provides recommendations or risk assessments to clinicians (or patients). It emphasizes not just static rules but possibly predictive algorithms. ONC now requires transparency about DSIs in certified systems[58][59].

- **EHR – Electronic Health Record:** Digital system for patient medical records and clinical workflow. EHR integration is key for AI deployment (delivering AI advice directly into clinician's ordering or documentation workflow).

- **EMA – European Medicines Agency:** The EU agency responsible for the scientific evaluation, supervision, and safety monitoring of medicines. EMA is paying attention to AI in drug development and in companion diagnostics, issuing guidance to align with the broader EU AI Act[25][26].

- **EU AI Act – European Union Artificial Intelligence Act:** Landmark EU legislation (entering into force 2024, compliance expected by 2025-26) to regulate AI systems by risk level. It classifies AI used in healthcare as "high-risk," imposing requirements for risk management, data quality, transparency, human oversight, etc., before such systems can be marketed in the EU[25][26].

- **FDA GMLP – Food and Drug Administration's Good Machine Learning Practice:** FDA's set of guiding principles (10 of them) for AI/ML in medical devices to ensure they are safe and effective[33][15]. Not formal regulations, but consensus guidelines shared by FDA, Health Canada, and UK MHRA.

- **FURM – Fair, Useful, Reliable, Measurable:** A framework from Stanford for evaluating AI models' readiness for deployment[12]. It corresponds to ensuring the model is fair (no undue bias), useful (addresses a real need & yields net benefit[5]), reliable (performs robustly and as expected), and measurable (impact can be tracked). Essentially a practical checklist similar to PACT-CARE's coverage.

- **FUTURE-AI:** Stands for Fairness, Universality, Traceability, Usability, Robustness, Explainability[9]. It's an international consensus guideline for trustworthy AI in healthcare (originating in EU for medical imaging AI, now broadly applicable)[63][10]. Each component is a principle to strive for in development and deployment.

- **HITL – Human-in-the-Loop:** A design where human(s) are actively involved in the AI system's operation – either in generating training data, validating outputs, or making final decisions. In our context, it specifically means a human oversees and can override AI decisions in clinical use.

- **HTI-1 – Health Data, Technology, and Interoperability ("HTI-1") Final Rule:** A 2023 rule by the US Office of the National Coordinator (ONC) updating EHR certification criteria. Among other things, it introduced **algorithmic transparency requirements for predictive decision support tools** in certified health IT[58][59]. It replaces older CDS criteria with new DSI (Decision Support Intervention) criteria, effective 2025.

- **LLM/LMM – Large Language Model / Large Multimodal Model:** Refers to very large machine learning models trained on extensive data (text for LLMs, or multiple data types like text+images for LMMs) that can perform a range of tasks (like GPT-4, etc.). In healthcare, they might be used for chatbot triage, medical Q&A, image+text analysis, etc. WHO's 2024 guidance addresses these specifically, urging caution and oversight due to their unpredictable outputs[71][62].

- **MDR/IVDR – Medical Device Regulation / In Vitro Diagnostic Regulation:** The EU regulations (in effect since 2021 for MDR and 2022 for IVDR) governing medical devices and diagnostics. These replaced older directives and impose stricter requirements. Software for medical purposes (including many AI) falls under MDR. Compliance involves classification, conformity assessment (often with notified bodies for higher risk classes), clinical evaluation, risk management, etc. AI-based diagnostics might be IVDR. Essentially, to deploy an AI in EU clinically, you likely need to meet MDR/IVDR unless it's exempt or not considered a medical device.

- **NNA – Number-Needed-to-Alert:** An adaptation of the Number-Needed-to-Treat concept. It asks: for each *alert or recommendation* an AI gives, how many such alerts result in one actual desired outcome? For instance, if 10 alerts yield 2 true positive actions that benefit patients, NNA = 5. It's a way to gauge how efficient an alerting system is for clinicians. Lower NNA is better (means most alerts are impactful). If NNA is high (like 50), it means a lot of noise per benefit, which could indicate low usefulness.

- **ONC – Office of the National Coordinator for Health IT:** The US federal entity that sets standards and certification for EHRs and health IT, among other coordination roles. They are the ones pushing for algorithmic transparency in EHRs (so clinicians know about the AI tools within) and interoperability. They don't "approve" AI like FDA, but they influence how AI gets integrated into mainstream health IT.

- **PCCP – Predetermined Change Control Plan:** FDA's framework (still being refined) for allowing AI device modifications post-approval. A PCCP is essentially a regulatory submission section where the developer pre-defines what changes the AI will undergo and how they will control them[17]. If FDA agrees, the device can then update within those bounds without a new submission every time. It's part of FDA's effort to regulate "adaptive" or continuously learning AI.

- **PCE – Pooled Cohort Equations:** The risk equations introduced in the 2013 ACC/AHA Guidelines to estimate 10-year risk of heart attack or stroke. It uses factors like age, sex, race, cholesterol, blood pressure, diabetes, smoking, etc. It's the basis for the ASCVD Risk Calculator widely used in preventive cardiology. Many AI attempts in cardiology aim to improve or recalibrate these equations with local data or additional factors[72][73].

- **SaMD – Software as a Medical Device:** A term from regulatory agencies like FDA/IMDRF referring to software intended to be used for medical purposes without being part of a hardware medical device. Many stand-alone AI apps are SaMD. SaMD still require regulatory clearance if they perform diagnoses, predictions, or treatment recommendations (unless they meet an exemption like some CDS). The term distinguishes pure software from software in a device.

- **WHO – World Health Organization:** A UN agency responsible for international public health. They don't regulate but provide guidance and coordinate global health efforts. Their reports on AI ethics (2021) and LMMs (2024) set important benchmarks for how health organizations and governments should approach AI[8][74]. Often local policies are influenced by WHO guidance.

---

By understanding and using these concepts and frameworks, one can better navigate the responsible development and deployment of AI in healthcare. PACT-CARE™ essentially merges these threads (clinical, technical, ethical, regulatory) into a single actionable framework – aiming to ensure that *AI in healthcare becomes not just smart, but also safe, fair, and truly beneficial in the hands of clinicians and for the good of patients*.

[1][5]

---


[1] External validation of the Epic sepsis predictive model in 2 county emergency departments - PubMed

https://pubmed.ncbi.nlm.nih.gov/39545248/

[2] [5] [6] [7] [11] [28] How Do We Ensure that Healthcare AI is Useful? | Stanford HAI

https://hai.stanford.edu/news/how-do-we-ensure-healthcare-ai-useful

[3] [4] Targeted Reminders Increase Prescriptions for High-Intensity Statins - American College of Cardiology

https://www.acc.org/About-ACC/Press-Releases/2023/03/05/14/27/Targeted-Reminders-Increase-Prescriptions

[8] [20] [21] [67] WHO outlines principles for ethics in health AI | The Verge

https://www.theverge.com/2021/6/30/22557119/who-ethics-ai-healthcare

[9] [10] [53] [63] The FUTURE-AI guidelines - 6 key principles of AI4HF project - AI4HF

https://www.ai4hf.com/news-updates/the-future-ai-guidelines-6-key-principles-of-ai4hf-project

[12] furm [Shah Lab]

https://shahlab.stanford.edu/doku.php?id=furm

[13] [14] [47] [2403.07911] Standing on FURM ground -- A framework for evaluating Fair, Useful, and Reliable AI Models in healthcare systems

https://arxiv.org/abs/2403.07911

[15] [33] [57] Good Machine Learning Practice for Medical Device Development: Guiding Principles | FDA

https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

[16] [17] Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles | FDA

https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles

[18] [19] [48] [58] [59] [60] [61] HHS, ONC HTI-1 Final Rule Introduces New Transparency Requirements for Artificial Intelligence in Certified Health IT | Mintz

https://www.mintz.com/insights-center/viewpoints/2146/2024-01-08-hhs-onc-hti-1-final-rule-introduces-new-transparency

[22] [23] [24] [62] [71] [74] WHO releases AI ethics and governance guidance for large multi-modal models

https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models

[25] [26] [27] [64] European Regulator Clarifies Guidance on the Use of AI in the Medicinal Product Lifecycle | GoodLifeSci

https://goodlifesci.sidley.com/2024/10/22/european-regulator-clarifies-guidance-on-the-use-of-ai-in-the-medicinal-product-lifecycle/

[29] [32] A simple, step-by-step guide to interpreting decision curve analysis

https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-019-0064-7

[30] Decision curve analysis for quantifying the additional benefit of a ...

https://www.fharrell.com/post/addmarkerdca/

[31] Optimizing Clinical Decision Making with Decision Curve Analysis

https://pmc.ncbi.nlm.nih.gov/articles/PMC10454914/

[34] [35] [36] [69] [70] Trust in healthcare AI must be felt by doctors and patients | World Economic Forum

https://www.weforum.org/stories/2025/08/healthcare-ai-trust/

[37] [38] [41] [42] [43] [44] [49] [50] [51] [52] [54] Study: AI Bolsters Sensitivity for Pneumothorax on CXR and Significantly Reduces Reporting Time

https://www.diagnosticimaging.com/view/study-ai-sensitivity-pneumothorax-cxr-significantly-reduces-reporting-time

[39] [40] Reducing Disparities in No Show Rates Using Predictive Model-Driven Live Appointment Reminders for At-Risk Patients: a Randomized Controlled Quality Improvement Initiative - PMC

https://pmc.ncbi.nlm.nih.gov/articles/PMC10150669/

[45] [46] Bankers: Clinical no-show reduction (Pitch) | BUSN39100 Augmented Intelligence

https://voices.uchicago.edu/201702busn3910001/2017/05/17/bankers-clinical-no-show-reduction-pitch/

[55] [56] Exploring the Role of Predictive Analytics in Mitigating No-Show Appointments and Enhancing Patient Engagement | Simbo AI - Blogs

https://www.simbo.ai/blog/exploring-the-role-of-predictive-analytics-in-mitigating-no-show-appointments-and-enhancing-patient-engagement-3640995/

[65] [66] Summary and assessment of EMA’s reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle :: Parexel

https://www.parexel.com/insights/blog/summary-and-assessment-of-emas-reflection-paper-on-the-use-of-artificial-intelligence-ai-in-the-medicinal-product-lifecycle

[68] Patients Don't Trust Physicians Who Use AI, Study Suggests

https://www.renalandurologynews.com/news/patients-dont-trust-physicians-who-use-ai/

[72] Automating and improving cardiovascular disease prediction using ...

https://www.sciencedirect.com/science/article/pii/S1386505622001009

[73] ASCVD Risk Estimator Plus - American College of Cardiology

https://www.acc.org/Tools-and-Practice-Support/Mobile-Resources/Features/2013-Prevention-Guidelines-ASCVD-Risk-Estimator