From AUC to Accountability: Metric Choices, Social Norms, and the Deployment of Therapeutic AI

David H. Silver
Rhea Labs (Rhea Fertility)
Singapore
david.silver@rhea-fertility.com
https://orcid.org/0000-0002-3071-304X

November 10, 2025

Abstract

Clinical and regulatory discussions about trustworthy therapeutic AI speak in ethical and legal terms, while technical work reports performance through AUC, F1, PPV, or survival concordance. The mapping between these numerical summaries and the social norms they support is implicit and often incoherent. This gap is acute for therapeutic AI systems built on spatially resolved data—such as digital pathology, radiology, and spatial omics—where models guide target selection, biomarker discovery, and responder prediction.

We provide a technical–normative analysis of common evaluation metrics. We formalize metric families by their aggregation operations: expectation-based (expected loss; AUC as a U-statistic over pairs), quantile/tail-based (median, upper quantiles, CVaR), supremum-type (worst-group risk, minimax regret), thresholded confusion-matrix ratios (PPV/precision, sensitivity, F1), and ranking metrics (top-k, average precision). For each family we identify the implicit social norm: maximizing average benefit, protecting typical patients, guarding against worst-case harms, or prioritizing top-predicted benefit.

We prove an incompatibility result showing that high AUC can coexist with very low worst-group sensitivity under deployment-relevant thresholding, and we illustrate further incompatibilities via clinical examples. We then propose a metric design framework: (i) explicit normative declarations, (ii) multi-objective evaluation with subgroup and tail-risk constraints, and (iii) deployment checklists tying thresholds to institutional responsibilities. The goal is not to replace ethical debate with formulas, but to make explicit and auditable the value judgments encoded by metric choices.

1 Introduction

Therapeutic AI systems support clinical decisions from biomarker discovery in digital pathology to stratifying treatment response in radiology and spatial omics. These systems are typically evaluated using area under the ROC curve (AUC), F1, positive predictive value (PPV), sensitivity/specificity, or survival concordance indices (C-index). The survival C-index is a pairwise concordance measure over comparable time-to-event pairs, with censoring-specific nuances such as comparable-pair definitions and inverse probability of censoring weighting (19; 3; 44).

Meanwhile, regulatory frameworks and clinical institutions discuss trustworthy AI in the language of ethics, accountability, and patient safety (16; 28; 11). The mapping between these two registers—scalar performance metrics and social norms—is usually left implicit. A system with AUC > 0.90 may be declared "clinically acceptable," yet the normative work such a threshold performs is rarely made explicit.

This gap creates predictable failures: systems with strong average metrics can exhibit unacceptable subgroup performance (35), violate implicit standards of care (45), or create liability ambiguity when failures occur (36). The problem is acute in therapeutic contexts, where spatially resolved data introduce heterogeneity and where deployment decisions directly affect treatment selection and patient stratification.

Our contribution. We formalize the normative implications of standard evaluation metrics for therapeutic AI. We show that metric families encode distinct social norms about acceptable risk, benefit distribution, and subgroup protection; we include a formal construction demonstrating incompatibility between high AUC and worst-group sensitivity at plausible deployment thresholds; and we propose a metric design and reporting framework that ties evaluation to institutional accountability.

Scope. We focus on supervised prediction in therapeutic contexts using spatial/high-content data. We briefly relate the survival C-index to pairwise concordance but do not develop censoring-specific theory. Our analysis applies when deployment affects individual care or population-level therapeutic strategies.

2 Metric Families: Formal Definitions

We formalize aggregation operations underlying common metrics and note their mathematical properties. Let $(X, Y, G) \sim P$ denote the population distribution over features, labels, and subgroup $G \in \mathcal{G} = \{G_1, \ldots, G_k\}$. Let $\widehat{P}$ denote the empirical distribution induced by a test dataset $\mathcal{D} = \{(x_i, y_i, g_i)\}_{i=1}^n$. Let $f : \mathcal{X} \to [0, 1]$ be a predictor and $\ell : [0, 1] \times \{0, 1\} \to \mathbb{R}_+$ a loss (e.g., absolute error or log loss). We use $\mathbb{E}_P[\cdot]$ and $\mathbb{E}_{\widehat{P}}[\cdot]$ to denote population and empirical expectations, respectively. We present population definitions for conceptual clarity and report empirical U-statistic/plug-in estimators in practice; unless noted, examples refer to empirical estimates.

2.1 Expectation-Based Aggregation

Expected loss (mean risk). The population and empirical risks are

$$R^{\text{pop}}_{\text{mean}}(f) = \mathbb{E}_P[\ell(f(X), Y)],$$
$$R^{\text{emp}}_{\text{mean}}(f) = \mathbb{E}_{\widehat{P}}[\ell(f(X), Y)].$$

Under 0-1 loss this is error rate; under log loss, cross-entropy.

AUC as a U-statistic over pairs. The population AUC can be written as

$$\text{AUC}(f) = \mathbb{P}_P\big[f(X^+) > f(X^-)\big] + \tfrac{1}{2}\mathbb{P}_P\big[f(X^+) = f(X^-)\big],$$

with independent draws $X^+ \sim P(\cdot \mid Y{=}1)$, $X^- \sim P(\cdot \mid Y{=}0)$. The empirical U-statistic estimator is

$$\widehat{\text{AUC}}(f) = \frac{1}{n_+ n_-} \sum_{i:Y_i=1} \sum_{j:Y_j=0} \Big(\mathbf{1}\{f(x_i) > f(x_j)\} + \tfrac{1}{2}\mathbf{1}\{f(x_i) = f(x_j)\}\Big),$$

where n_+ and n_- are the numbers of positives and negatives. AUC is non-decomposable over i.i.d. examples even though it is expectation-based (17). For survival analysis, the C-index analogously computes concordance over comparable pairs with corrections for censoring (19; 3; 44).

Subgroup decomposition. Let $\text{AUC}(f \mid g, h)$ denote the concordance for positive examples from group g versus negative examples from group h. Then

$$\text{AUC}(f) = \sum_{g,h \in \mathcal{G}} w_{g,h}\, \text{AUC}(f \mid g, h),$$
$$w_{g,h} = \frac{\mathbb{P}(G{=}g, Y{=}1)\, \mathbb{P}(G{=}h, Y{=}0)}{\mathbb{P}(Y{=}1)\mathbb{P}(Y{=}0)}.$$

Weights depend on subgroup prevalences $\mathbb{P}(G{=}g, Y{=}y)$.

Calibration error. Idealized integrated calibration error (ICE) with L_1 norm is

$$\text{ICE}(f) = \mathbb{E}_P\big[\big|f(X) - \mathbb{E}_P[Y \mid f(X)]\big|\big].$$

Empirical estimates such as expected calibration error (ECE) and its adaptive variants are binned plug-in estimators; their magnitudes depend on binning and can be biased (often downward), so values are not directly comparable across binning schemes (15; 34). Post-hoc methods (isotonic regression, beta calibration) estimate a calibration mapping on held-out data; they are not themselves calibration error estimators (30; 25). The Brier score $\mathbb{E}[(Y - f(X))^2]$ admits a calibration–refinement decomposition (29).

2.2 Quantile and Tail-Risk Aggregation

Quantiles. For $\alpha \in (0, 1)$, the α-quantile of the loss $\ell(f(X), Y)$ under P is

$$Q_\alpha(f) = \inf\{t \in \mathbb{R} : \mathbb{P}_P[\ell(f(X), Y) \leq t] \geq \alpha\}.$$

Quantiles (VaR) are not coherent risk measures (they fail subadditivity) and are non-convex in distributions, which affects robustness and constrained optimization.

CVaR. Conditional Value-at-Risk at level α is the expected loss in the $(1{-}\alpha)$ upper tail; we adopt the Rockafellar–Uryasev formulation (39):

$$\text{CVaR}_\alpha(f) = \min_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{1 - \alpha}\, \mathbb{E}_P\big[(\ell(f(X), Y) - \tau)_+\big] \right\},$$

where $(z)_+ = \max(0, z)$. CVaR is coherent and convex.

2.3 Supremum-Type Aggregation

Worst-group risk and minimax regret. For subgroups $\mathcal{G}$,

$$R_{\text{sup}}(f) = \max_{g \in \mathcal{G}}\, \mathbb{E}_{P(\cdot|G=g)}[\ell(f(X), Y)].$$

Minimax regret compares to the subgroup-optimal predictor in a hypothesis class $\mathcal{H}$:

$$R_{\text{regret}}(f) = \max_{g \in \mathcal{G}} \Big(\mathbb{E}_{P(\cdot|G=g)}[\ell(f(X), Y)] - \inf_{f' \in \mathcal{H}} \mathbb{E}_{P(\cdot|G=g)}[\ell(f'(X), Y)]$$

Both R_{sup} and CVaR are coherent risk measures (monotone, subadditive, positive homogeneous, translation invariant). This mathematical structure aligns with "safety-first" reasoning; normative guarantees (e.g., anti-discrimination) require that subgroups align with protected classes and that constraints be enforced.

2.4 Thresholded Confusion-Matrix Ratios

Ratios such as PPV (precision), sensitivity (TPR), specificity (TNR), F1, and Jaccard are functions of confusion-matrix counts at a threshold and are non-decomposable; some are conditional means (e.g., TPR is an average over $Y{=}1$) but all are threshold-dependent and induce trade-offs (31; 32). PPV's dependence on prevalence is salient clinically. For completeness, $\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$; under certain score-distribution assumptions, the F1-optimal threshold approximately equalizes precision and recall.

Family	Aggregation	Examples
Expectation	$\mathbb{E}[\cdot]$	Error, cross-entropy; AUC (pairwise U-stat; non-decomp.)
Quantile/tail	Q_α, CVaR	Median, 95th pct., CVaR
Supremum	$\max_{g \in \mathcal{G}} \mathbb{E}[\cdot \mid G{=}g]$	Worst-group risk, regret
Conf.-ratio	Thresholded ratios	PPV, TPR/TNR, F1, Jaccard
Ranking	Weighted by rank	Precision@k, AP

Table 1: Metric families and aggregation operations.

2.5 Ranking-Based Aggregation

Top-k and average precision (AP). Let $(x_{(i)}, y_{(i)})$ be the pair corresponding to the ith highest score. Then

$$\text{Precision@}k = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}[y_{(i)} = 1], \quad \text{AP} = \frac{1}{\text{Pos}} \sum_{i:\, y_{(i)}=1} \text{Precision@}i,$$

where Pos is the number of positives (9). In practice, ties are handled by consistent tie-breaking or credit sharing; implementation choices can affect AP.

Summary and bridge. Table 1 summarizes these families. We now use these aggregation operators as the organizing principle for explicating the social norms each metric family operationalizes.

3 Implicit Social Norms

Each metric family encodes an implicit social norm about risk distribution and benefit allocation.

3.1 Expectation-Based Metrics: Average Benefit

Minimizing $\mathbb{E}[\ell]$ or maximizing AUC operationalizes a utilitarian norm: maximize aggregate benefit. AUC is invariant to any strictly monotone transform of scores; thus it ignores absolute risk calibration and any loss asymmetries tied to clinical thresholds. Its threshold-agnostic nature can conflict with context-specific decision thresholds and asymmetric costs (46). For survival tasks, the C-index shares the pairwise concordance view over comparable pairs, and—like AUC—can be insensitive to systematic subgroup mis-ranking (19; 44).

3.2 Quantile/Tail Metrics: Protect the Typical

Minimizing the median or constraining upper-tail quantiles encodes: the typical patient (or, e.g., 95%) should receive acceptable care, even if a small tail experiences higher loss. Quantile constraints protect most patients but can still sacrifice small, identifiable minorities; for anti-discrimination aims, supremum-type constraints are more appropriate. CVaR strengthens tail protection by focusing on the severity of worst cases.

3.3 Supremum Metrics: Guard the Worst-Off

Worst-group or minimax criteria encode: every identifiable subgroup must meet a minimum standard; no group may be systematically disadvantaged. CVaR and worst-group risk are coherent risk measures (4), aligning with safety-first norms; when subgroups align with protected classes and constraints are enforced, they operationalize anti-discrimination requirements.

3.4 Thresholded Ratios: Operating-Point Guarantees

Confusion-matrix ratios (PPV, TPR, F1) encode: performance at a chosen operating threshold matters, with explicit precision–recall trade-offs. Optimizing F1 can increase false positives in low-prevalence settings and may be misaligned with clinical utility (22). Decision curve analysis links a risk-threshold to utilities via net benefit, mapping TPR and FPR to clinical value at that threshold (46).

3.5 Ranking Metrics: Prioritize Scarce Benefits

Maximizing Precision@k or AP encodes: correctly identifying top-ranked cases is prioritized (e.g., scarce trial slots), even if lower-ranked performance is modest.

4 Incompatibilities with Clinical and Legal Norms

We formalize and illustrate incompatibilities between standard metrics and explicit clinical/legal expectations. In deployment, a single global operating threshold (and abstention/deferral policy) is typically used across groups; incompatibilities should be demonstrated under such rules.

4.1 A Formal Incompatibility

Proposition 1 (High AUC with low worst-group sensitivity)**.** *Fix* $\delta \in (0,1)$ *and* $\varepsilon \in (0,1)$*. There exists a binary classification problem with two groups* $G \in \{A, B\}$*,* $\mathbb{P}(G{=}B) = \delta$*, equal class prevalence across groups* $\mathbb{P}(Y{=}1 \mid G{=}A) = \mathbb{P}(Y{=}1 \mid G{=}B) \in (0,1)$*, and a scoring function* f *such that* $\text{AUC}(f) \geq 1 - \delta$ *(indeed* $1 - \delta + \frac{1}{2}\delta^2$*) while the sensitivity (TPR) for group* B *at some fixed global deployment threshold satisfies* $\text{TPR}_B \leq \varepsilon$*.*

Proof sketch. AUC decomposition. Let $\text{AUC}(f \mid g, h)$ be the concordance for positives from g versus negatives from h. Then

$$\text{AUC}(f) = \sum_{g,h} w_{g,h}\, \text{AUC}(f \mid g, h),$$

$$w_{g,h} = \frac{\mathbb{P}(G{=}g, Y{=}1)\mathbb{P}(G{=}h, Y{=}0)}{\mathbb{P}(Y{=}1)\mathbb{P}(Y{=}0)}.$$

Under equal class prevalence across groups, $w_{A,A} = (1-\delta)^2$, $w_{A,B} = (1-\delta)\delta$, $w_{B,A} = \delta(1-\delta)$, and $w_{B,B} = \delta^2$.

Score construction. Let group A be perfectly separated: $f(X) \in [2,3]$ for $Y{=}1$ and $f(X) \in [1,2)$ for $Y{=}0$. Let group B be uninformative: $f(X) \sim U[0,1]$ independent of Y. Then $\mathrm{AUC}(f \mid A, A) = 1$, $\mathrm{AUC}(f \mid A, B) = 1$ (any A positive scores above any B negative), $\mathrm{AUC}(f \mid B, A) = 0$ (any B positive scores below any A negative), and $\mathrm{AUC}(f \mid B, B) = 0.5$. Therefore

$$\begin{aligned}\mathrm{AUC}(f) &= (1-\delta)^2 \cdot 1 + (1-\delta)\delta \cdot 1 + \delta(1-\delta) \cdot 0 + \delta^2 \cdot 0.5 \\ &= 1 - \delta + \tfrac{1}{2}\delta^2 \geq 1 - \delta.\end{aligned}$$

Global threshold and group-B TPR. For any threshold $\tau \in [0,1]$, $\mathrm{TPR}_B = \mathbb{P}[U(0,1) > \tau] = 1 - \tau$. Choosing a single global threshold $\tau = 1-\varepsilon$ yields $\mathrm{TPR}_B = \varepsilon$. For clinically plausible global thresholds maximizing discrimination on pooled data (see Corollary 1), we obtain $\mathrm{TPR}_B = 0$. □

Corollary 1 (Plausible global selection rules)**.** *In the setting of Proposition 1, any single global threshold* $\tau^\star$ *that maximizes Youden's* $J = \mathrm{TPR} - \mathrm{FPR}$ *on pooled data satisfies* $\tau^\star \in [2,3]$*, yielding* $\mathrm{TPR}_B = 0$*. Likewise, any rule that sets a global threshold to achieve an overall PPV target or overall* $\mathrm{FPR} \leq \eta$ *for some* $\eta < 1 - \delta$ *selects* $\tau^\star \geq 2$*, implying* $\mathrm{TPR}_B = 0$*.*

Proof sketch. For $\tau \in [2,3]$, $\mathrm{TPR}_A = 1$, $\mathrm{FPR}_A = 0$, and $\mathrm{TPR}_B = \mathrm{FPR}_B = 0$; hence $J = 1 - \delta$, which is maximal compared to $\tau < 2$ (where $J \leq 1-\delta$) or $\tau < 1$ (where $J = 0$ due to equal increases in pooled TPR and FPR). Any overall PPV or FPR target below that achievable with $\tau < 2$ requires $\tau \geq 2$, which forces $\mathrm{TPR}_B = 0$. □

The construction shows that very high AUC can coexist with vanishing subgroup sensitivity under deployment-relevant, single-threshold rules.

4.2 Minimum Standard Violations

Clinical norm. No patient should fall below a minimum standard given the care context.

Metric conflict. Expectation-based metrics permit arbitrarily poor performance on subsets if the average remains high.

Example 1 (Spatial pathology bias)**.** *A tumor detection model performs with sensitivity* 0.98 *on 90% well-stained slides but only* 0.40 *on 10% poorly stained slides from under-resourced sites; the average sensitivity is* 0.922*. This appears strong, yet 10% of patients receive systematically substandard care. Overall AUC can remain high while worst-group sensitivity is low; subgroup AUC and worst-group constraints are needed (17; 40). A single threshold chosen on pooled validation to meet a global sensitivity or PPV target can mask these gaps because mixture weighting over-represents the majority site's distribution. Empirically, stain and site variability are well documented (43; 5).*

Legal implication. Consistent failures in identifiable contexts (e.g., poor staining) can constitute negligence regardless of strong averages (14; 27).

4.3 Subgroup Fairness Failures

Clinical norm. Identifiable demographic or clinical subgroups should not be systematically disadvantaged solely due to algorithmic design (28; 6; 37).

Metric conflict. AUC or F1 can remain strong despite large disparities across subgroups.

Example 2 (Responder prediction heterogeneity (PPV disparity))**.** *Suppose 1000 patients: 700 White, 300 Black. At a single deployment threshold, the model predicts positive for 400 White patients with 300 true positives (PPV*$_W{=}0.75$*) and for 200 Black patients with 90 true positives (PPV*$_B{=}0.45$*). Overall PPV is* $390/600 = 0.65$*. This strong overall PPV masks a large subgroup disparity. PPV depends on base rates; disparities can arise from true prevalence differences even when TPR/FPR are equal. Report both threshold-invariant metrics (e.g., AUC, subgroup AUC) and threshold-dependent metrics (e.g., TPR, PPV) by subgroup.*

Training under imbalance can worsen minority performance if minority data are noisier or underrepresented (35; 20).

4.4 High Average, Catastrophic Tail

Clinical norm. Rare but severe failures can be unacceptable if preventable (16).

Metric conflict. Mean-based metrics are insensitive to tails. Consider ℓ with $\ell{=}0$ for correct care, $\ell{=}10$ for missing aggressive cancers. Model f_1: $\mathbb{P}[\ell{=}10] = 0.5\%$; model f_2: $\mathbb{P}[\ell{=}10] = 2.5\%$; both tuned so $\mathbb{E}[\ell] = 0.25$. Mean risk is identical, but $\mathrm{CVaR}_{0.95}(f_2) > \mathrm{CVaR}_{0.95}(f_1)$, revealing the clinically worse tail (39).

4.5 Systematic Bias in Ranking

With equal prevalence and similar average TPR, under-representation at top-k can arise from score distribution shifts or miscalibration.

Example 3 (Ranking bias in treatment prioritization)**.** *A model ranks candidates for a scarce trial; groups* A *and* B *have equal prevalence. The top-100 list includes 80 from A and 20 from B. High AUC is compatible with such disparity when scores are miscalibrated or variances differ across groups. Under exchangeability (e.g., if, conditional on* Y*, score distributions are identical across groups), the expected top-*k *composition is proportional to group share; deviations then suggest potential allocation disparity. We recommend cautious two-sample proportion tests, complemented by calibration and subgroup audits (41).*

5 Framework for Normative Alignment

5.1 Explicit Normative Declaration

Pair each reported metric with the norm it operationalizes:

- AUC: average discriminative ability across pairwise comparisons; threshold-agnostic; not a clinical utility.
- Worst-group error: ensures no subgroup exceeds error threshold ϵ.
- Precision@k: prioritizes correct identification of top k patients for scarce interventions.
- Decision curves: link thresholds to utilities via net benefit, relating TPR/FPR to net clinical benefit at a specified risk threshold (46).

5.2 Multi-Objective Evaluation

Definition 1 (Constrained evaluation)**.** *A model f is clinically acceptable if, under a single global operating point (and a single abstention/deferral policy) applied across all subgroups,*

$$\mathbb{E}_P[\ell(f(X), Y)] \leq \tau_{\text{avg}}, \tag{1}$$

$$\max_{g \in \mathcal{G}} \mathbb{E}_{P(\cdot|G=g)}[\ell(f(X), Y)] \leq \tau_{\text{group}}, \tag{2}$$

$$Q_{0.95}(f) \leq \tau_{\text{tail}}, \quad \text{CVaR}_{0.95}(f) \leq \tau_{\text{cvar}}. \tag{3}$$

Setting $(\tau_{\text{avg}}, \tau_{\text{group}}, \tau_{\text{tail}}, \tau_{\text{cvar}})$ requires deliberation among clinicians, patients, ethicists, and regulators, and should be documented in the governance plan.

Implementation (concise guide).

- Audit. Pre-specify subgroups (including salient intersections) and primary/secondary endpoints. Enforce minimum N for reporting; suppress or flag metrics below this N.
- Training. If constraints fail, retrain under group-robust/tail-robust objectives (Group DRO; CVaR minimization) (40; 47); reductions approaches and subgroup fairness auditing can help (1; 23).
- Statistics. Report uncertainty via DeLong intervals/tests for AUCs and subgroup AUC comparisons (10); Wilson or Newcombe intervals for proportions (Clopper–Pearson is conservative) (48; 33); use stratified bootstrap for complex metrics; control multiplicity in subgroup audits.
- Safeguards. Implement abstention/deferral triggered by uncertainty or OOD scores (e.g., calibrated confidence thresholds, conformal p-values (2)). Set thresholds to satisfy subgroup constraints on the non-abstained set; log deferrals and outcomes for monitoring.

5.3 Deployment Checklists and Responsibilities

- Override and escalation. If subgroup sensitivity falls below a clinical threshold (e.g., < 0.80), deploy only with mandatory human review for affected subgroups and documented mitigation plans.
- Liability allocation. Deploying with known subgroup failures shifts responsibility to the institution; strong average AUC does not excuse foreseeable subgroup harms (36; 27; 14). Regulatory frameworks (e.g., FDA/IMDRF GMLP for SaMD; EU MDR) locate responsibility for known failure modes and mandate postmarket surveillance (21; 13).
- Monitoring and drift. Continuously monitor group-specific and intersectional metrics with predefined triggers for retraining or rollback; publish periodic reports.

6 Discussion

Expectation-based metrics (AUC, cross-entropy) dominate because they enable scalable stochastic optimization; quantile, supremum, and non-decomposable objectives are harder to optimize and audit (40; 31). Yet spatial/high-content modalities (whole-slide pathology, radiology, spatial transcriptomics) exhibit preparation artifacts, site shifts, anatomical variability, and rare morphologies. Average metrics can obscure these tails; subgroup and tail-risk constraints surface them and better align with safety requirements (43; 5). Ethical frameworks are not interchangeable: utilitarian, maximin, and allocation-focused norms correspond to distinct metric families (expectation, suprema/CVaR, rankings). Scalarizing objectives can aid optimization, but for governance it risks hiding trade-offs (24; 7; 18).

7 Conclusion

Standard evaluation metrics for therapeutic AI encode specific social norms about acceptable risk, benefit distribution, and subgroup protection. These implicit norms can conflict with clinical and legal expectations that no patient falls below minimum standards, high-risk subgroups receive protection, and systematic bias is unacceptable regardless of averages. Our incompatibility construction (Proposition 1) illustrates why average-discrimination metrics alone are insufficient; explicit subgroup and tail-risk constraints are necessary for accountable deployment.

We proposed an auditable framework that combines explicit normative declarations with multi-objective evaluation (average, subgroup, and tail constraints including quantiles and CVaR) and deployment checklists tying thresholds to institutional responsibilities. The goal is not to replace ethical debate with formulas, but to ensure that what is measured matches what clinical practice claims to value.

References

[1] Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69.

[2] Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511.*

[3] Antolini, L.; Boracchi, P.; and Biganzoli, E. 2005. A time-dependent discrimination index for survival data. *Statistics in Medicine* 24(24):3927–3944.

[4] Artzner, P.; Delbaen, F.; Eber, J.-M.; and Heath, D. 1999. Coherent measures of risk. *Mathematical Finance* 9(3):203–228.

[5] Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Silva, V. W. K.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* 25(8):1301–1309.

[6] Char, D. S.; Shah, N. H.; and Magnus, D. 2018. Implementing Machine Learning in Health Care—Addressing Ethical Challenges. *New England Journal of Medicine* 378(11):981–983.

[7] Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2):153–163.

[8] Clopper, C. J.; and Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26(4):404–413.

[9] Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of ICML*, 233–240.

[10] DeLong, E. R.; DeLong, D. M.; and Clarke-Pearson, D. L. 1988. Comparing the areas under two or more correlated ROC curves: a nonparametric approach. *Biometrics* 44(3):837–845.

[11] Drabiak, K. 2022. Leveraging law and ethics to promote safe and reliable AI/ML in healthcare. *Frontiers in Nuclear Medicine* 2:983340.

[12] Fordham, D. E.; Rosentraub, D.; Polsky, A. L.; Aviram, T.; Wolf, Y.; Perl, O.; Devir, A.; Rosentraub, S.; Silver, D. H.; and Zamir, Y. G. 2022. Embryologist agreement when assessing blastocyst implantation probability: is data-driven prediction the solution to embryo assessment subjectivity? *Human Reproduction* 37(10):2275–2289.

[13] European Union. 2017. Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices (EU MDR). *Official Journal of the European Union.*

[14] Gerke, S.; Minssen, T.; and Cohen, G. 2020. Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare*, 295–336.

[15] Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *ICML*, 1321–1330.

[16] Habli, I.; Lawton, T.; and Porter, Z. 2020. Artificial intelligence in health care: accountability and safety. *Bulletin of the World Health Organization* 98(4):251–259.

[17] Hand, D. J. 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* 77:103–123.

[18] Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NeurIPS*.

[19] Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; and Rosati, R. A. 1982. Evaluating the yield of medical tests. *JAMA* 247(18):2543–2546.

[20] Hashimoto, T. B.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 1929–1938.

[21] International Medical Device Regulators Forum (IMDRF). 2021. Good Machine Learning Practice for Medical Device Development: Guiding Principles. *IMDRF Publication IMDRF/SaMD WG/N78.*

[22] Jansche, M. 2005. Maximum expected F-measure training of logistic regression models. In *Proceedings of HLT/EMNLP*, 692–699.

[23] Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *NeurIPS*.

[24] Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807.*

[25] Kull, M.; Silva Filho, T. M.; and Flach, P. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *AISTATS*.

[26] Kumar, A.; Liang, P.; and Ma, T. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*.

[27] Lawton, T.; Morgan, P.; Porter, Z.; Hickey, S.; Cunningham, A.; Hughes, N.; Iacovides, I.; Jia, Y.; Sharma, V.; and Habli, I. 2024. Clinicians risk becoming 'liability sinks' for artificial intelligence. *Future Healthcare Journal* 11(1):100007.

[28] Morley, J.; Machado, C.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; and Floridi, L. 2020. The ethics of AI in health care: A mapping review. *Social Science & Medicine* 260:113172.

[29] Murphy, A. H. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* 12(4):595–600.

[30] Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI*.

[31] Narasimhan, H.; Vaida, S.; and Agarwal, S. 2014. A structural SVM based approach for optimizing F-measures. In *Advances in Neural Information Processing Systems*.

[32] Narasimhan, H.; Kar, P.; and Jain, P. 2015. Optimizing non-decomposable measures with a surrogate loss. In *International Conference on Machine Learning*.

[33] Newcombe, R. G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 17(8):857–872.

[34] Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring calibration in deep learning. *arXiv:1904.01685*.

[35] Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453.

[36] Price, W. N.; and Cohen, I. G. 2023. Locating Liability for Medical AI. *Social Science Research Network*.

[37] Rajkomar, A.; Hardt, M.; and Howell, M. D. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine* 169(12):866–872.

[38] Rave, G.; Fordham, D. E.; Bronstein, A. M.; and Silver, D. H. 2024. Enhancing Predictive Accuracy in Embryo Implantation: The Bonna Algorithm and its Clinical Implications. In *Lecture Notes in Computer Science*, volume 14806, 161–175.

[39] Rockafellar, R. T.; and Uryasev, S. 2000. Optimization of conditional value-at-risk. *Journal of Risk* 2(3):21–41.

[40] Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-group performance. In *International Conference on Learning Representations*.

[41] Singh, A.; and Joachims, T. 2018. Fairness of exposure in rankings. In *Proceedings of KDD*.

[42] Silver, D. H.; Feder, M.; Ben-Meir, A.; Pasternak, Y.; Drukker, L.; Elad, D.; Simchen, M. J.; and Reichman, O. 2020. Data-Driven Prediction of Embryo Implantation Probability Using IVF Time-lapse Imaging. In *Medical Imaging with Deep Learning*.

[43] Tellez, D.; Litjens, G.; Bándi, P.; Bulten, W.; Bokhorst, J.-M.; Ciompi, F.; and van der Laak, J. 2019. Quantifying the effects of data augmentation and stain color normalization in computational pathology. *Medical Image Analysis* 58:101544.

[44] Uno, H.; Cai, T.; Pencina, M. J.; D'Agostino, R. B.; and Wei, L.-J. 2011. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* 30(10):1105–1117.

[45] Verdicchio, M.; and Perin, A. 2022. When Doctors and AI Interact: on Human Responsibility for Artificial Risks. *Philosophy & Technology* 35:71.

[46] Vickers, A. J.; and Elkin, E. B. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 26(6):565–574.

[47] Williamson, R. C.; and Menon, A. K. 2019. Fairness risk measures. In *International Conference on Machine Learning*, 6786–6797.

[48] Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22(158):209–212.