ellis | INSTITUTE FINLAND

Aalto-yliopisto

HEALTHTECH FINLAND · SUSTAINABILITY WORKING GROUP

Reliable and Generalizable AI Systems

Faithfulness, robustness and fairness for health technology

Qi Chen

PI, ELLIS Institute Finland · Assistant Professor, Aalto University

1 September

AI Already Creates Real Value in Healthcare

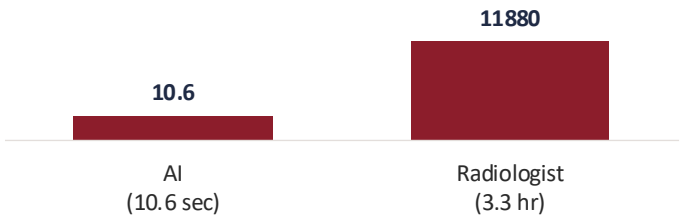
Diagnostic support — flagging findings in imaging, labs, and records for clinician review.

Workflow — prioritizing cases and reducing time-to-treatment.

Discovery — accelerating drug and biomarker discovery from limited experimental data.

Rib-fracture AI, ED (23,251 films)
10.6 sec read time vs. 3.3 hrs (P<.001)

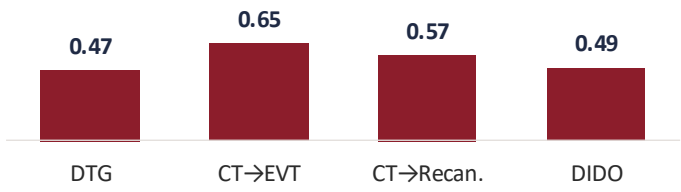
Median read time (log scale, sec)



Huang et al., JMIR Med Inform, 2026 (MacKay Memorial Hospital)

Viz.ai LVO triage, meta-analysis (n=15,595)
Faster CT-to-treatment & door-to-groin time (p<.001)

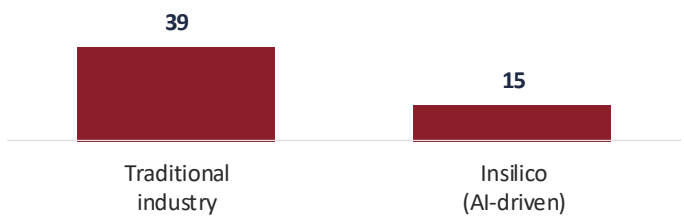
Workflow time reduction (SMD magnitude, favors AI)



Sarhan et al., meta-analysis, Int J Stroke, 2025 (n=15,595)

Rentosertib (AI-discovered TNIK inhibitor)
First AI-designed drug to reach Phase III

Time to preclinical candidate (months)



Insilico Medicine; Nature Medicine, 2025

Three Questions Behind “Reliable and Generalizable AI4Health”

1

Faithfulness

“Does the output reflect what's actually true — not just what looks plausible?”

2

Robustness

“Does the model still work under distribution shifts?”

3

Fairness

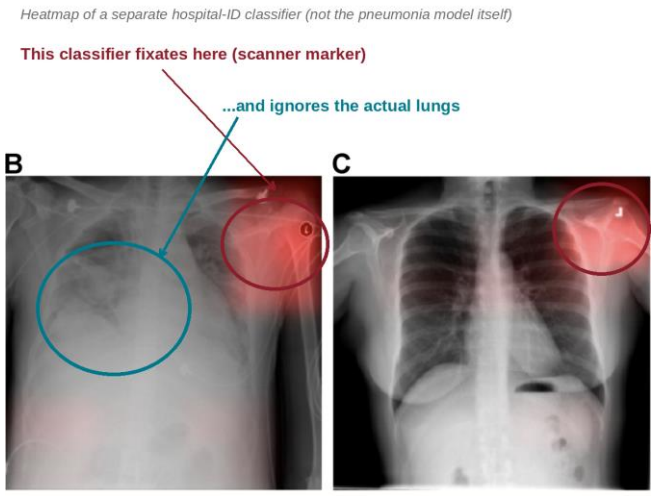
“Does it work equally well for every subgroup?”

Each is a different way an AI system can quietly stop being trustworthy, and each has concrete, testable practices behind it.

Faithfulness: Plausible Is Not Enough

Plausible outputs can mislead clinical decisions when they misrepresent the patient's data or the model's reasoning.

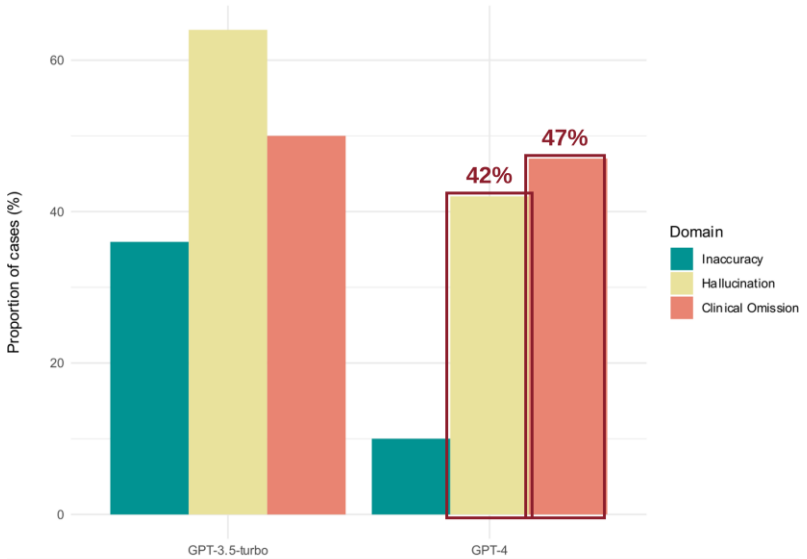
1 · An explanation hides an unreliable prediction



What fails: a classifier trained only to spot hospital system hits 99.95% accuracy using this marker — evidence the pneumonia model likely exploits the same shortcut.

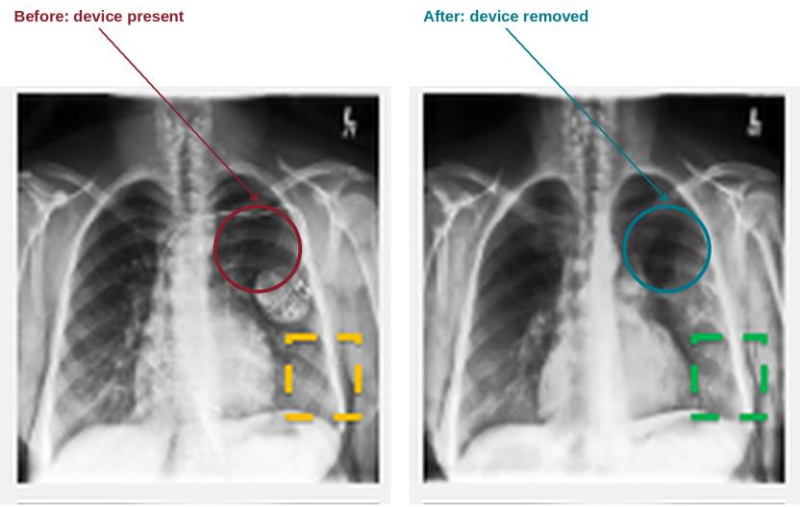
Potential consequence: clinicians may trust the pneumonia prediction without recognizing it relies on an unreliable shortcut.

2 · A generated summary changes the patient's information



What fails: GPT-4 still hallucinated 42% of summaries and omitted 47% of details, including allergies.
Potential consequence: an unverified error like this could contribute to an unsafe medication choice.

3 · An edited scan misrepresents disease progression



What fails (if unfaithful): overreach would alter the anatomy shown above too; under-editing would leave the device in place.

Potential consequence: a clinician comparing scans could misread a fabricated or unresolved finding.

(1) Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. "Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study." PLOS Medicine, 2018. (Actual figure.)

(2) Williams CYK, Bains J, Tang T, Patel K, Lucas AN, Chen F, Miao BY, Butte AJ, Kornblith AE. "Evaluating Large Language Models for Drafting Emergency Department Encounter Summaries." PLOS Digital Health, 2025. (Actual figure; allergy scenario is illustrative.)

Faithfulness in Generative Editing

Two Opposite Failure Modes:

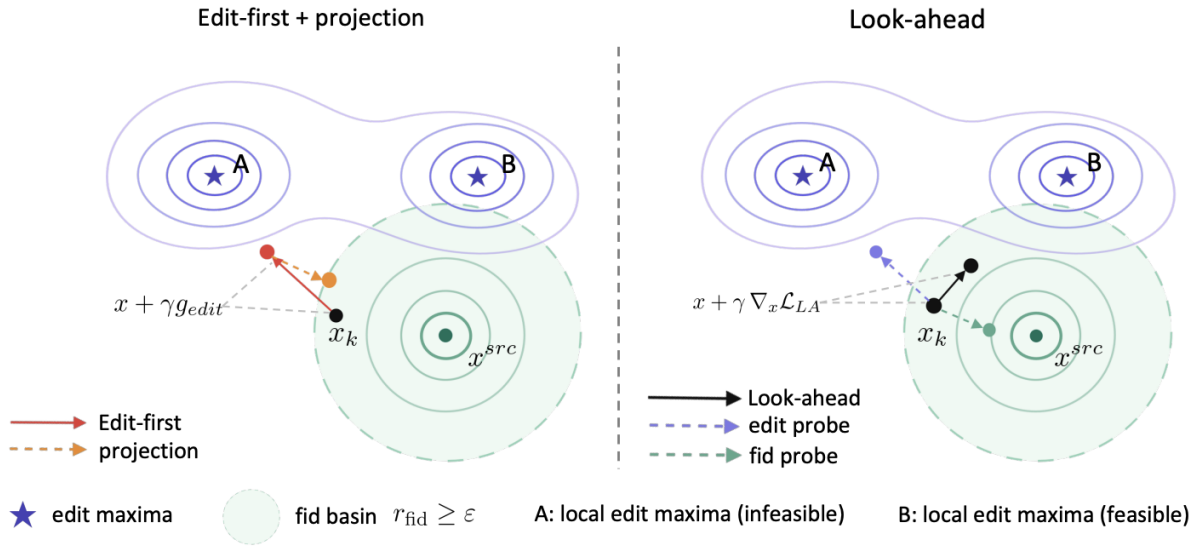
O Overreach

A model changes more than it was asked to: drifting past the target edit into anatomy, style, or details that should have stayed untouched.

U Under-editing

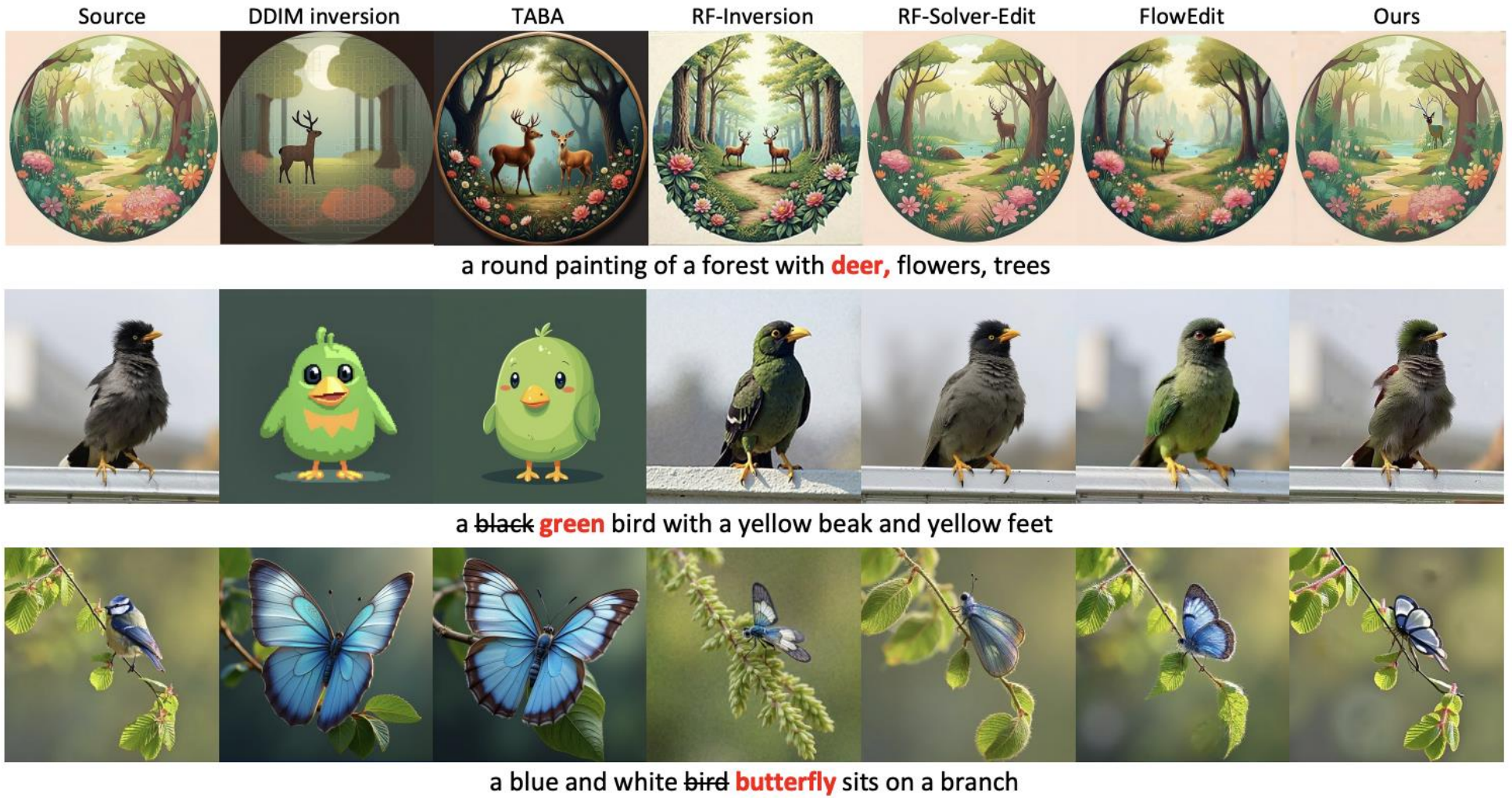
The model plays it too safe: the target condition isn't actually reflected in the output, so the counterfactual doesn't test what it was meant to test.

How to improve faithfulness for both objectives?



Doudou Zhang & Qi Chen, "Constrained Look-Ahead Guidance for Inference-Aware Flow Editing," under review, 2026

Improved Faithfulness



Doudou Zhang & Qi Chen, "Constrained Look-Ahead Guidance for Inference-Aware Flow Editing," under review, 2026

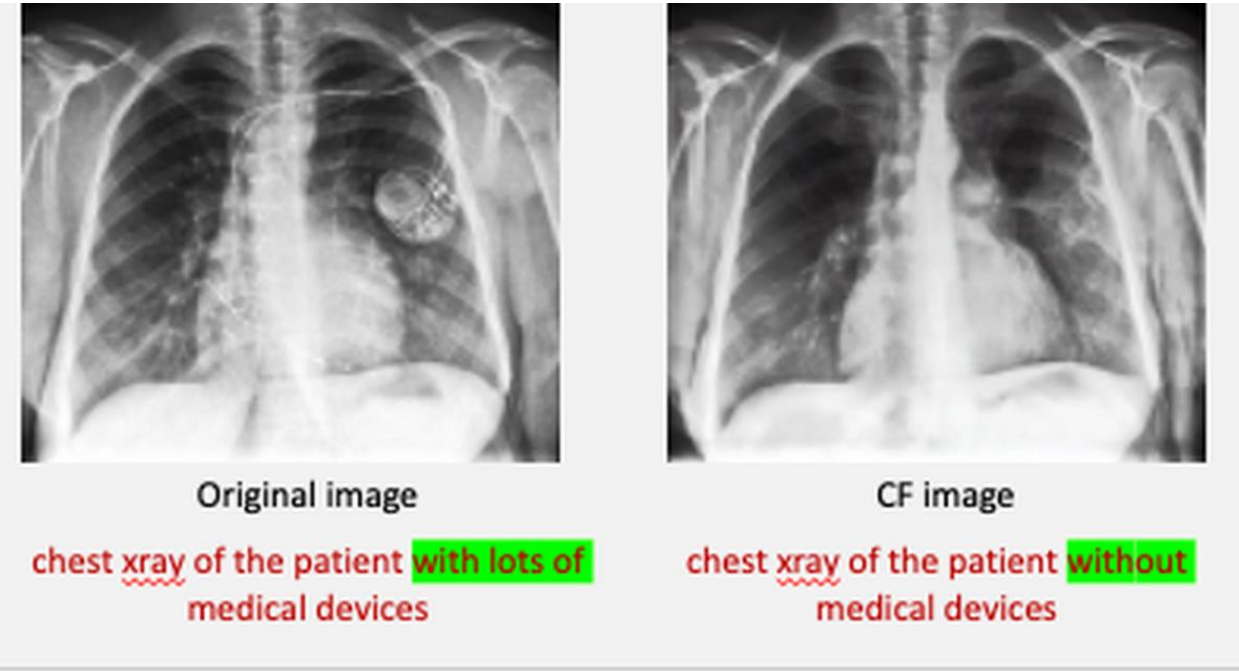
Revisit Counterfactual X-ray Editing

Task

Edit a chest X-ray to reflect a target clinical condition while preserving patient-specific anatomy.

Why it matters

Enables counterfactual (CF) reasoning, stress-testing, and model evaluation, and supports data augmentation.

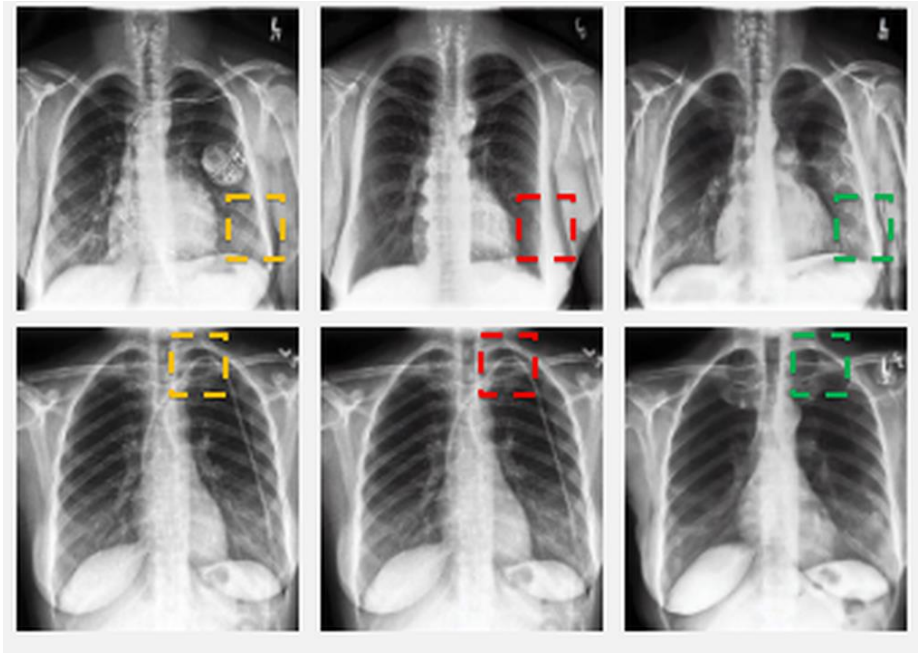


"Chest X-ray of the patient with lots of medical devices" → "without medical devices"

Doudou Zhang & Qi Chen, "Constrained Look-Ahead Guidance for Inference-Aware Flow Editing," under review, 2026

Checking Fidelity at Every Step, Not After the Fact

Existing methods either invert the image first (slow, and often loses fidelity) or edit freely and hope the result still looks like the patient. Our method checks an explicit fidelity constraint at every denoising step: so the edit can't drift into an anatomically implausible result before it's caught.



Original *CF (baseline)* *CF (ours)*

Our edits are more targeted and preserve more of the patient's anatomy.

Method	L_1 ↓	CPG ↑	Success ↑
Baseline (PRISM)	0.122	0.061	50.98%
Ours	**0.055**	**0.125**	**70.59%**

Evaluated on 99 held-out CheXpert chest X-rays.

- L_1: image preservation.
- CPG: counterfactual prediction-probability gain.
- Success: edits that flip the target label while staying inside the fidelity constraint.

The takeaway for deployment: an explicit, checkable constraint is what turns a plausible-looking edit into a reliable one.

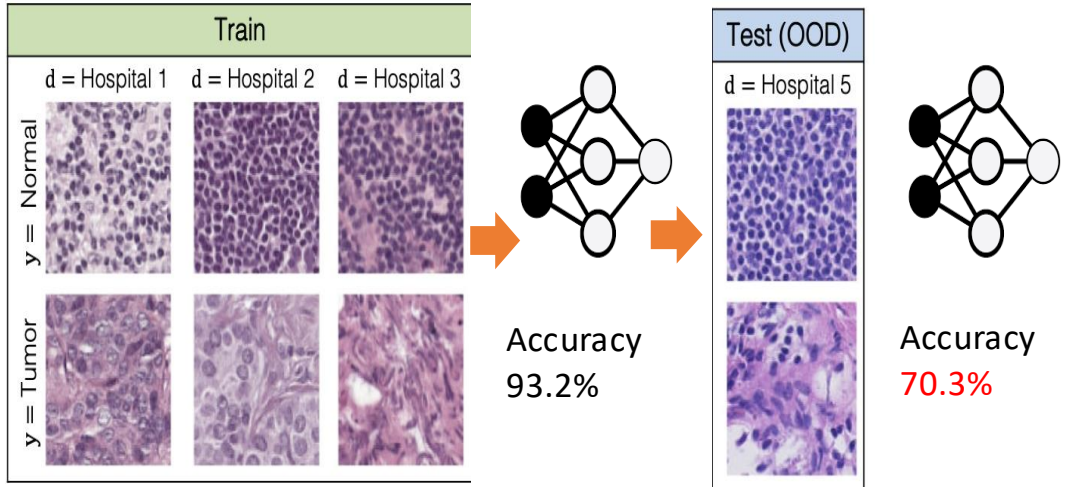
Robustness: Will It Still Work under Distribution Shifts?

A model validated at one hospital (with its particular scanners and patient population) faces a much broader test when deployed elsewhere.

Hospital A (training site)

→ *deployed at Hospital B (different scanner, demographics, protocols)*

Validated performance drops — without any change to the model itself.

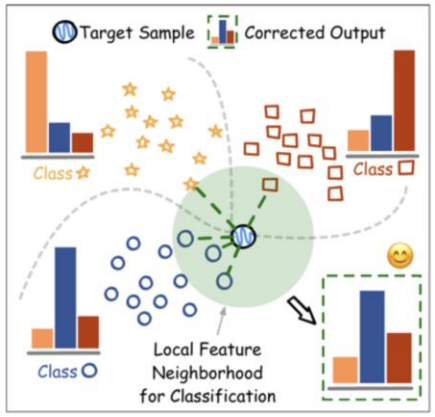


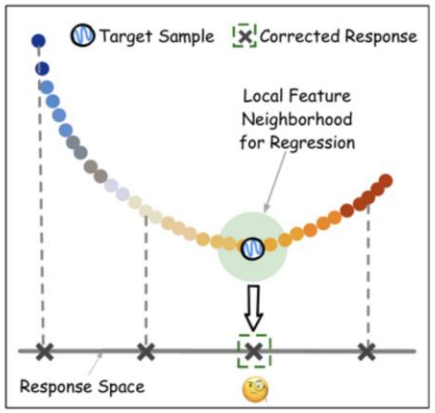
Robustness under distribution shift is a general challenge that extends beyond healthcare. Our prior work offers several approaches to addressing it:

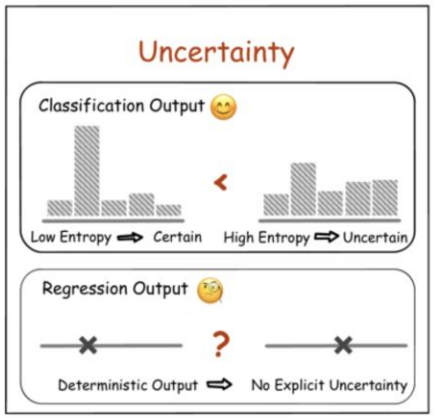
Qi Chen*, and Mario Marchand✉. Algorithm-Dependent Bounds for Representation Learning of Multi-Source Domain Adaptation. AISTATS 2023.
Changjian Shui, **Qi Chen**, Jun Wen, Fan Zhou, Christian Gagné, and Boyu Wang. A novel domain adaptation theory with Jensen–Shannon divergence. KBS 2022.

Harder Cases in Health Data

- Continuous clinical value
- No source-data for privacy concerns



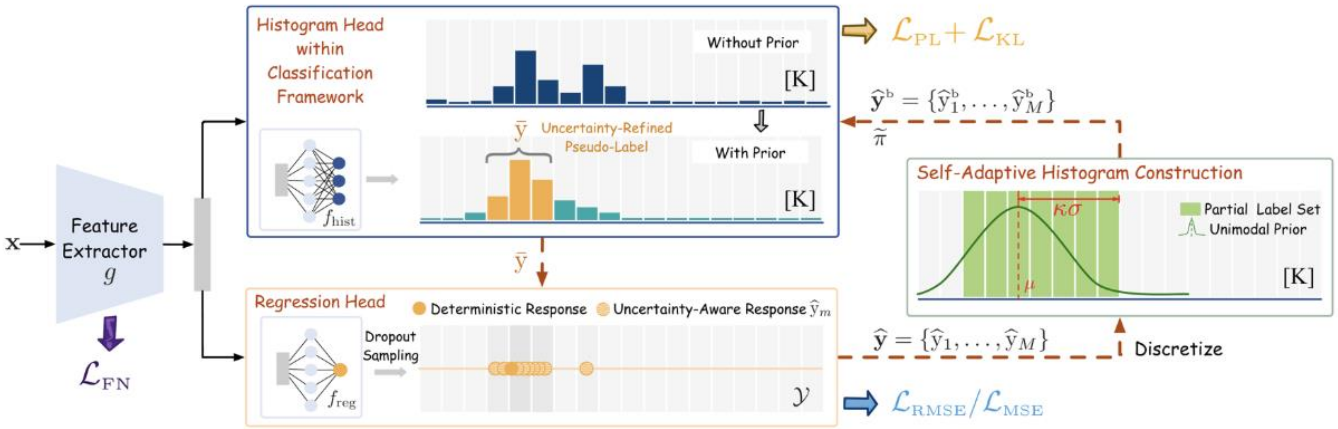




Why regression makes this harder

- **Classification SFDA** has two helpful signals:
  - clustered feature space for neighborhood-based correction;
  - softmax distribution for output entropy (confidence) estimation.
- **Regression SFDA** lacks both:
  - smooth feature manifold gives limited local correction signals;
  - deterministic scalar output gives no explicit uncertainty info.

Giving a Regression Model a Sense of Its Own Uncertainty



- **Histogram learning**

$$\mathcal{L}_{\mathrm{hist}}=\lambda_{\mathrm{PL}}\mathcal{L}_{\mathrm{PL}}+\lambda_{\mathrm{prior}}\mathcal{L}_{\mathrm{KL}}$$

- **Regression adaptation**

$$\mathcal{L}_{\mathrm{reg}}=\lambda_{\mathrm{MSE}}\mathcal{L}_{\mathrm{MSE}}\ \text{or}\ \lambda_{\mathrm{RMSE}}\mathcal{L}_{\mathrm{RMSE}}$$

- **Feature consistency**

$$\mathcal{L}_{\mathrm{FN}}\triangleq\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left|\|g(\mathrm{x}_i^{\mathrm{T}})\|-\|g_{\mathrm{S}}(\mathrm{x}_i^{\mathrm{T}})\|\right|$$

Age estimation, adapting across a domain shift (MAE, lower is better)

Method	Uses source data?	MAE ↓
No adaptation	—	6.55
Best prior source-free method	No	6.23
Ours (MERCI)	**No**	**5.83**

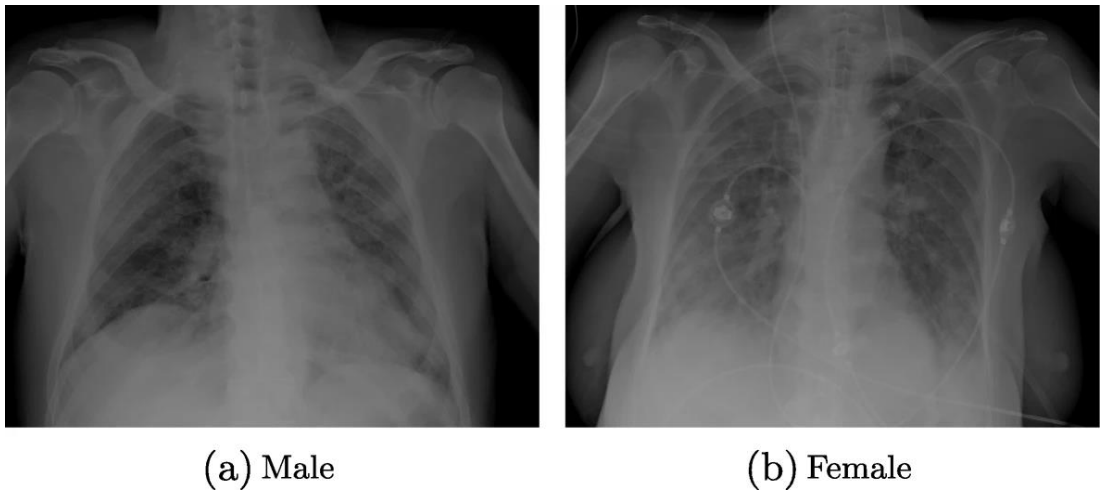
UTKFace age-estimation benchmark, adapting across a demographic split, no source data or target labels used.

The benchmarks in the paper are age, head-pose, and price prediction (not clinical data), but the same constraint (no source data, no target labels) is exactly what a real cross-hospital deployment faces.

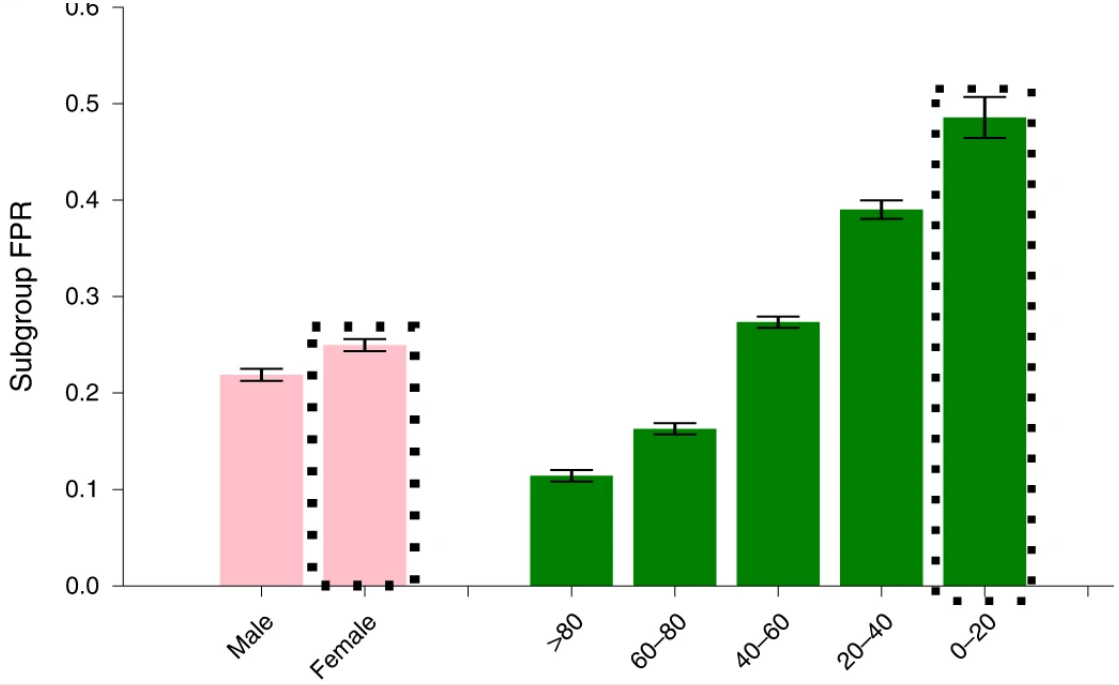
Gezheng Xu, Qi Chen, Qihao Zeng, Charles Ling & Boyu Wang, "Discretized Density-Guided Source-Free Domain Adaptation for Regression," ICML 2026 Spotlight

Fairness: Faithful and Robust Isn't Enough

A model can be faithful and robust, but still be systematically wrong for specific patient groups.



(a) Male (b) Female



Chest radiographs across subgroups (left) and the resulting underdiagnosis false-positive rate by sex and age group (right).

$$E[Y \mid f(x) = \tau, A = \text{Black patients}] > E[Y \mid f(x) = \tau, A = \text{White patients}]$$

Actual sickness — Clinic-risk score

"…At a given clinic-risk score, Black patients are considerably sicker than White patients…"

Seyyed-Kalantari et al., Nature Medicine, 2021

Obermeyer et al, "Dissecting racial bias in an algorithm used to manage the health of populations." Science, 2019

Improve Fairness through Algorithmic Design

$$\text{Suf}_{\text{Gap}}(f) \leq \frac{4}{|A|}\sum_a E_x[|f - E[Y|X,A]||A=a]$$

Fairness metric

Subgroup Bayes model (optimal in subgroup A) f_a

$$\min_{Q\in\mathcal{Q}} \frac{1}{|\mathcal{A}|}\sum_a \text{KL}(\overline{Q}_a^\star \| Q)$$ (Upper-level)

Randomized predictor

$$\text{s.t. } \overline{Q}_a^\star = \text{argmin}_{Q_a\in\mathcal{Q}}\{\mathbb{E}_{\tilde{f}_a\sim Q_a}\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_a) + \lambda \text{KL}(Q_a\|Q)\}, \forall a\in\mathcal{A}$$ (Lower-level)

Generalization error through PAC-Bayes

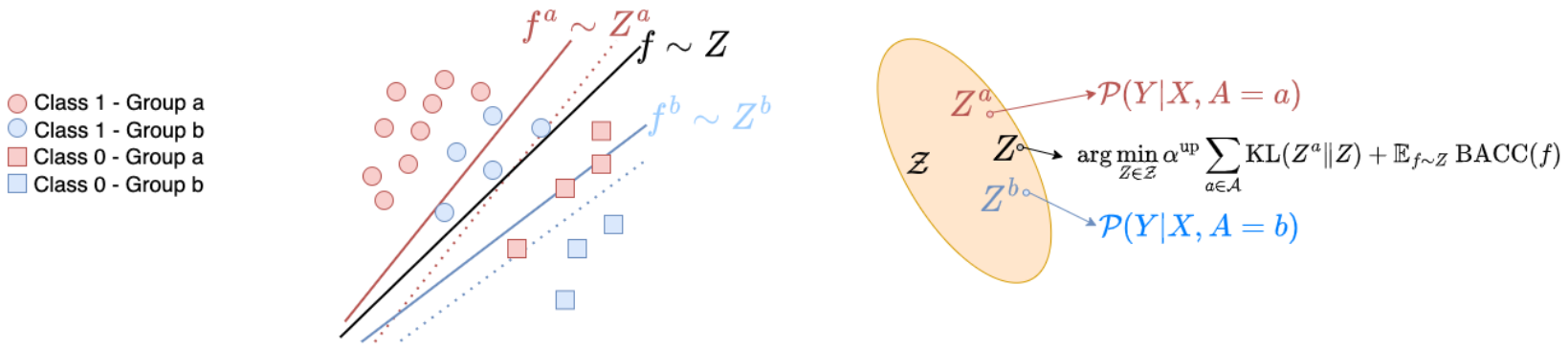
$$\frac{1}{|A|}\sum_a E_{f_a\sim Q_a} L_a(f_a) \leq \frac{1}{|A|}\sum_a E_{f_a\sim Q_a}\hat{L}_a(f_a) + \frac{C}{\sqrt{|A|m}}\sum_a \sqrt{\text{KL}(Q_a\|Q)} + O\left(\sqrt{\frac{1}{|A|m}}\right)$$

Test error | Training error | Informative Prior

"On learning fairness and accuracy on multiple subgroups." [Changjian Shui, Gezheng Xu, **Qi Chen et al.**, NeurIPS'22].

Application in Health Data

Task: distinguishing Alzheimer's disease (AD) from mild cognitive impairment (MCI) using **ADNI data** (5,137 samples, cognitive scores, MRI region volumes, and PET/CSF biomarkers, split into four race-based subgroups where over 90% of patients are White).



Method	Balanced accuracy ↑	Sufficiency gap ↓
ERM	86.87%	0.1701
FAMS — our NeurIPS'22	**83.69%**	**0.1328**
FACIMS — follow-up, UAI'23	88.97%	0.1052

Tarzanagh et al., *"Fairness-Aware Class Imbalanced Learning on Multiple Subgroups"* (UAI 2023, method: FACIMS)

THE COMMON THREAD

One Question, Three Disguises

Faithfulness, robustness, and fairness are really the same underlying question, asked from three angles:

“Will this model's behavior stay grounded and generalize beyond the exact patients, sites, and examples it was validated on?”

Why This Is Harder in Health Tech

Getting faithfulness, robustness, and fairness right is a data problem as much as a modeling one, and health data is uniquely constrained.

Scarce

Labeled clinical data is limited by cohort size, rare conditions, and annotation cost.

Costly

Every additional case often means new ethics approval, expert time, or trial cost.

Sensitive

Sharing or pooling data across institutions is constrained by privacy and governance rules.

A system that isn't faithful, robust, or fair typically also needs more data to fix after the fact. Building these properties in from the start is inseparable from being data-efficient.

Training-cost and data-scaling trends: Cottier et al., 2024; Sorsch et al., NeurIPS 2022

What This Means for HealthTech Companies

1. Validate across sites, scanners, and subgroups, not just one held-out split
2. Validate generated outputs against source data and task constraints before use; faithfulness is not guaranteed.
3. Monitor deployed models for drift after every update, not just at launch
4. Choose fairness criteria with clinical and regulatory stakeholders, for the specific context
5. Invest in data efficiency and reuse, since quality matters as much as volume

Where We Could Collaborate

The purpose of today is a shared discussion and hopefully some concrete next steps.

Evaluation methodology

Joint frameworks for testing faithfulness, robustness, and subgroup fairness before deployment.

Case studies

Applying theory and methods from Aalto/ELLIS research to a company's specific data, models, or deployment setting.

People

PhD and postdoc collaborations, or shared projects between member companies and Aalto.

Thank You

Discussion welcome.