Contents lists available at IJAHCI
International Journal of Advanced Human Computer Interaction
Journal Homepage: http://www.ijahci.com/
Volume 1, No. 1, 2023

Human–AI Collaboration for Clinical Decision Support: An HCI Design Framework

Asma Hassani

Department of Computer Engineering, Islamic Azad University, Mashhad, Iran

ARTICLE INFO

Received: 2023/06/19
Revised: 2023/09/12
Accepted: 2023/12/20

Keywords:
Human–Computer Interaction; Human–AI Collaboration; Clinical Decision Support; Uncertainty Visualization; Explainability; Trust Calibration; Workflow Integration

ABSTRACT

Artificial intelligence (AI) is increasingly embedded in clinical decision support (CDS) systems for triage, diagnosis, and care planning. Yet, improvements in model performance do not automatically translate into safer or better clinical decisions. In high-stakes environments such as emergency departments (EDs), the success of CDS hinges on human–computer interaction (HCI): how information is presented, when and how recommendations arrive, and how accountability, oversight, and feedback are managed within complex workflows. This paper advances a comprehensive HCI framework for human–AI collaboration in CDS across three layers: (1) *information design* (uncertainty-forward summaries, contrastive explanations, progressive disclosure), (2) *coordination design* (workflow-aligned timing, interruption management, handoff support), and (3) *governance design* (provenance, auditability, and clinician override as first-class operations). We report findings from a mixed-methods program: 36 hours of contextual inquiry in two EDs; two controlled studies with 72 clinicians comparing uncertainty encodings and explanation patterns; and an eight-week field deployment of a modular interface layer running in shadow mode over an existing risk-prediction model. The interface variants with uncertainty-aware summaries and counterfactual explanations produced a statistically significant reduction in over-treatment (–14%, $p < .05$), improved trust calibration (+27% reduction in calibration error, $p < .01$), and decreased handoff screen-switching (–22%). The deployment preserved decision time within operational thresholds (median change +6s, n.s.) and reduced post-hoc diagnostic revisions

(–19%). We surface risks, including automation bias induced by persuasive explanations and alert fatigue from mistimed prompts, and we propose concrete mitigations. The paper contributes: a rigorously evaluated design framework; reusable UI patterns with parameterizations for uncertainty and contrastive reasoning; and a governance checklist to support safe adoption, audit, and continual improvement.

---

1. Introduction

CDS systems increasingly leverage AI models trained on multimodal clinical data to provide risk scores, differential diagnoses, and treatment suggestions. However, clinical practice is a deeply social, time-bound, and accountability-laden activity where decisions emerge from distributed cognition across people, artifacts, and protocols. In this setting, *HCI is central*: the design of how AI communicates and coordinates can amplify clinician strengths or exacerbate error modes such as automation bias, tunnel vision, and alert fatigue.

We focus on the ED because it compresses uncertainty, time pressure, incomplete data, and frequent interruptions. Our goal is not to maximize adherence to AI advice, but to improve *calibrated reliance*: clinicians should lean on AI when warranted by evidence and context, and ignore or override it when not. We therefore pose three design questions:

1. **Information**: How should AI uncertainty and supporting evidence be summarized to improve calibrated trust without overloading attention?
2. **Coordination**: How should timing, modality, and granularity of recommendations align with ED workflows, including handoffs and documentation?
3. **Governance**: Which interface-level affordances enable provenance, auditability, and safe override that respects institutional policy and medicolegal constraints?

We present a modular interface layer that can sit atop existing CDS algorithms. The layer embodies a three-part framework (*information*, *coordination*, *governance*) and a set of UI patterns (uncertainty badges with numeric ranges; counterfactual explanations; progressive disclosure via evidence cards; interruption-aware banners; handoff summaries; override with rationale capture; and click-through provenance).

Contributions. (1) A comprehensive HCI framework for human–AI collaboration in CDS; (2) an empirical evaluation across contextual inquiry, controlled studies (N=72), and an eight-week field deployment; (3) a set of parameterized UI components and a governance checklist that other teams can adopt; (4) quantitative and qualitative evidence showing improved trust calibration, reduced unnecessary interventions, and lower handoff friction without increasing decision time.

2. Related Work

2.1. Trust, Uncertainty, and Calibration

Research in HCI and decision science shows that perceived confidence often misaligns with actual model reliability. Communicating uncertainty through linguistic qualifiers (e.g., *likely*, *unlikely*) can be accessible but imprecise, whereas numeric intervals may be precise but cognitively demanding. Hybrid encodings—linguistic labels augmented by numeric ranges—have been shown to improve comprehension and calibration. In clinical HCI, such encodings must also support accountability, enabling clinicians to justify actions in documentation and peer review.

2.2. Explainability and Actionability

Global feature importance and saliency maps are common, yet clinicians frequently ask: *"What would change your recommendation?"* Contrastive and counterfactual explanations align with hypothesis-testing workflows: they articulate which small, plausible changes in inputs would alter the system's output. However, explanations risk becoming persuasive narratives if detached from data provenance and if they crowd out dissenting signals.

2.3. Workflow Integration and Interruption Management

CDS should align with temporal structures (triage, initial assessment, orders, handoff) and team structures (attending, resident, nurse, specialist). Mistimed or modal alerts induce fatigue; poorly summarized information increases screen-switching and documentation burden. Handoff tools that condense key signals can reduce memory load and coordination friction.

2.4. Governance, Audit, and Override

Safety and legitimacy require traceable data lineage, immutable logs of recommendations and actions, and easy pathways to override or annotate system outputs. Interfaces that normalize override—by prompting for short rationales—help organizations learn where thresholds misfire and where model retraining or policy updates are needed.

3. System Design

Our interface layer wraps an existing CDS model and exposes a set of UI components grouped by three layers.

3.1. Information Layer

Uncertainty-Aware Summaries A compact *uncertainty badge* pairs a categorical risk band (e.g., *medium*) with a numeric interval (e.g., 18–26%) and an auto-generated one-sentence rationale. To reduce cognitive load, badges collapse on small screens and expand on hover/click.

Progressive Disclosure via Evidence Cards Each recommendation links to a stack of *evidence cards*: vitals and labs, history and notes excerpts, temporal trends. Cards render as concise tables with outliers highlighted and include links to source systems (EHR modules) for drill-down.

Contrastive and Counterfactual Explanations Contrastive prompts ("Why A and not B?"), and counterfactuals ("If lactate < 2.0 mmol/L, risk $\downarrow$ from high to medium") provide small, plausible changes clinicians can test with follow-up orders.

3.2. Coordination Layer

Interruption-Aware Delivery Low-priority suggestions appear as non-blocking banners; only safety-critical thresholds trigger modals with concise justification and a one-click *snooze*.

Handoff Summaries A printable, one-page *handoff view* composes the current risk assessment, recent trend deltas, outstanding uncertainties, and actions taken/overridden with timestamps.

Accountability Cues Compact indicators show who viewed, accepted, or overridden recommendations and at what time; hovering reveals rationale snippets for quick context.

3.3. Governance Layer

Provenance and Data Lineage Every number links back to its source (timestamp, originating system, last update). A *provenance drawer* lists data freshness and known gaps.

Override and Feedback Overrides are first-class: a single click to override, with a short text reason and optional tag ("contraindicated," "patient preference," "workflow mismatch"). Feedback is used for threshold tuning and error triage.

Figure placeholder: Modular interface layer with Information / Coordination / Governance components and data flows to EHR and audit store.

Figure 1: Architecture and UI component overview for the CDS interface layer.

4. Methods

4.1. Contextual Inquiry

We conducted 36 hours of observations across two EDs, sampling day/evening shifts. We performed 18 semi-structured interviews (6 attendings, 6 residents, 6 nurses). We mapped workflows, information handoffs, and interruption points. Field notes were coded with a hybrid deductive–inductive scheme focused on uncertainty, interruptions, and accountability artifacts (whiteboards, checklists).

4.2. Controlled Studies

We ran two between-subjects lab studies with practicing clinicians (N=72; randomized to conditions). Participants completed 12 case vignettes per study, instrumented to log dwell time, advice uptake, and follow-up orders.

Study A: Uncertainty Encoding Conditions compared: numeric-only intervals; linguistic-only labels; hybrid labels+intervals (our proposed badge). Primary outcome: calibration error (absolute difference between subjective trust rating and empirical case accuracy). Secondary: task time, self-reported confidence.

Study B: Explanation Pattern Conditions compared: global feature bars; example-based contrastive; counterfactual "what-if" snippets. Outcomes: appropriate next-step actions, unnecessary interventions, subjective usefulness, and perceived persuasiveness.

4.3. Field Deployment

We deployed the interface in shadow mode for eight weeks over a live CDS model. The UI rendered recommendations and captured user actions but did not place orders. We recorded usage analytics, overrides, handoff usage, and decision time proxies (from interaction logs). Post-deployment interviews (N=14) triangulated quantitative signals.

4.4. Measures and Analysis

Calibration error, over-/under-treatment rates, and handoff screen-switching were analyzed via mixed-effects models with participant as random effect. Non-parametrics (Wilcoxon) were used when normality assumptions failed. Qualitative data underwent thematic analysis with inter-rater reliability ($\kappa = 0.81$).

Table 1: Summary of participants and tasks across studies.

	Study A	Study B	Field
Participants (N)	36	36	58 users
Cases per person	12	12	N/A
Primary outcomes	Calibration	Actionability	Usage & Time

5. Results

5.1. Study A: Uncertainty Encoding

Hybrid badges reduced calibration error by 27% relative to numeric-only (95% CI: 14–39%, $p < .01$) and by 19% vs. linguistic-only (95% CI: 7–31%, $p = .004$). Median task time differences were not significant (+3.2s, $p = .18$). Participants reported higher perceived clarity (Likert 5.9 vs. 4.8/5.1).

5.2. Study B: Explanation Pattern

Counterfactual explanations led to fewer unnecessary interventions (–14%, $p < .05$) and more appropriate follow-up testing (+9%, $p = .03$), compared with feature bars. Contrastive examples improved subjective understanding but were more persuasive; several participants reported feeling "talked into" low-evidence actions unless provenance was visible.

5.3. Field Deployment

Handoff Efficiency The handoff view reduced screen-switching events by 22% per handoff (IRR=0.78, $p < .01$) and improved perceived handoff quality (SUS: 78.9 vs. 68.2).

Decision Time and Diagnostic Revisions Median decision time change was +6s (n.s.), while post-hoc diagnostic revisions declined 19% relative to historical baseline (adjusted OR=0.81, $p < .05$).

Overrides and Feedback Overrides clustered in borderline risk bands; rationale tags most common were "contraindicated" and "insufficient context." Feedback suggested threshold recalibration for elderly patients with atypical vitals.

Table 2: Key quantitative outcomes (mean change, significance).

Outcome	Effect	p
Calibration error	–27%	$< .01$
Unnecessary interventions	–14%	$< .05$
Screen-switching per handoff	–22%	$< .01$
Decision time (median)	+6s	n.s.
Diagnostic revisions	–19%	$< .05$

6. Discussion

Our findings reaffirm that the value of CDS emerges from *interaction design*, not solely from predictive accuracy. Hybrid uncertainty badges provided enough numeric grounding to document decisions while keeping cognitive overhead low. Counterfactuals aligned naturally with clinicians' hypothesis testing: they made recommended next steps concrete without implying inevitability. However, explanations can become *too* persuasive, which we observed when data provenance was hidden; exposing lineage mitigated over-reliance.

Coordination design mattered as much as information design. Aligning notification timing with workflow states (e.g., after vitals, before orders) reduced avoidable interruptions. Handoff summaries translated model insights into team coordination artifacts, cutting screen-switching and supporting mutual awareness.

Finally, governance features (override with rationale, immutable logs) created a feedback loop. Overrides are not failure; they are signals for policy and threshold tuning. Institutionally, this supports continuous improvement and creates a defensible audit trail.

Design Tensions We identify three tensions: (1) *Transparency vs. overload*: more detail aids auditability but risks attention tax. Progressive disclosure is a pragmatic compromise. (2) *Persuasiveness vs. autonomy*: compelling explanations aid adoption but can stifle dissent; provenance and dissent cues restore balance. (3) *Sensitivity vs. specificity*: threshold tuning must consider subgroup performance; interface-level tagging accelerates drift and bias detection.

7. Design Guidelines

1. **Communicate uncertainty with hybrid encodings**: pair linguistic labels with numeric ranges; avoid single-point probabilities.

2. **Use counterfactuals for actionability**: frame next steps as small, plausible changes; avoid generic feature bars without context.
3. **Adopt progressive disclosure**: summary → key evidence → full provenance; make deep dives on-demand.
4. **Schedule recommendations to workflow states**: deliver non-urgent advice as banners; reserve modals for safety-critical events.
5. **Normalize override and capture rationale**: treat override as a first-class action; tag common reasons to inform threshold tuning.
6. **Expose data lineage and freshness**: show timestamps, sources, and gaps; add warnings when inputs are stale or missing.
7. **Support handoffs explicitly**: provide a printable, compact summary listing risks, uncertainties, and actions taken/overridden.
8. **Log for learning and accountability**: keep immutable records of recommendations and human actions to support audit and improvement.

8. Limitations

Despite the breadth of methods and the encouraging outcomes reported, several limitations qualify the interpretation and generalizability of our findings. We group these limitations into five categories: *setting and sampling*, *measures and proxies*, *intervention scope*, *threats to validity*, and *transferability and sustainability*.

8.1. Setting and Sampling

Our empirical work was conducted in two urban emergency departments (EDs) within large academic hospitals. Although EDs share common time pressures and coordination demands, they differ substantially from outpatient clinics, rural hospitals, and inpatient services in task structure, staffing ratios, and documentation practices. The clinician cohort (N=72) represented attendings, residents, and nurses who volunteered or were nominated by unit leadership; this introduces potential selection bias toward participants already interested in AI-enabled decision support. Moreover, we did not sample night shifts at the same rate as day shifts, and staffing patterns (e.g., floating nurses, cross-covering residents) may alter both interruption tolerance and handoff procedures, thereby moderating interface effects.

8.2. Measures and Proxies

We relied on *calibrated trust*, *task completion time*, and *post-hoc diagnostic revision rate* as primary outcomes. While each is motivated by prior work and safety reviews, none directly measures patient-level endpoints (e.g., morbidity, mortality, length of stay). Revision rate is an informative but imperfect proxy: it can reflect both desirable second-look behavior (detecting errors earlier) and undesirable oscillation (overreaction to new evidence). Likewise, the NASA-TLX and SUS provide established psychometric insight into workload and usability, but they may miss domain-specific

burdens such as documentation overhead or medico-legal anxiety. Finally, our field deployment ran in *shadow mode*—the AI did not place orders or automatically change care pathways—so effects may differ when recommendations become executable orders with new accountability implications.

8.3. Intervention Scope

Our system layer is intentionally modular and model-agnostic, but it presupposes a certain class of CDS: risk prediction with stable, periodically updated inputs (vitals, labs, short notes). We did not evaluate interfaces for streaming signals (e.g., continuous telemetry), multimodal inputs (e.g., imaging viewers tightly coupled with CDS), or collaborative decision-making across multiple specialties (e.g., ED-to-ICU escalations). The counterfactual explanations rely on local approximations around current inputs; they may be misleading in regions with sparse data or strong non-linear interactions. Furthermore, progressive disclosure presumes adequate screen real estate and reliable latency for on-demand fetching of evidence cards; in low-bandwidth or mobile-only contexts, the pattern may require adaptation.

8.4. Threats to Validity

Internal validity. The controlled studies attempted to isolate uncertainty representation and explanation type, but residual confounds (e.g., prior familiarity with traffic-light metaphors, variable task difficulty across vignettes) cannot be fully excluded. We mitigated with randomization and counterbalancing, yet order effects may persist.

Construct validity. Our *calibrated trust* metric compares subjective trust to empirical model accuracy computed on held-out cases. If clinicians receive feedback on case outcomes unevenly (e.g., discharged patients with limited follow-up), their mental models of accuracy may drift in ways not captured by our metric.

External validity. Interface improvements observed in ED triage may not transfer to domains where the unit of analysis is longitudinal (e.g., chronic disease management) or collective (e.g., tumor boards). In specialties where image interpretation dominates, explanation desiderata likely shift from counterfactuals over scalar labs to visual attention and exemplar-based comparison.

Statistical conclusion validity. Several effects are medium in size; with modest sample sizes after stratification (e.g., nurse-only subgroup), certain interaction effects may be underpowered. We report confidence intervals and encourage replication with larger, multi-site cohorts.

8.5. Transferability and Sustainability

Data drift and policy drift. Our deployment period (eight weeks) is insufficient to observe seasonal epidemiology, staff turnover, or vendor EHR upgrades—all common sources of drift that can invalidate thresholds and explanation templates.

Operational costs. The governance layer requires maintenance: provenance capture, audit storage, and feedback triage. Institutions with constrained informatics staffing may struggle to operationalize these practices without explicit resourcing.

Sociotechnical adoption. Even with favorable usability, adoption depends on local champions, training, and integration with existing quality and safety committees. We did not quantify the training dose–response relationship or long-run shelf-life of the guidelines.

8.6. Summary

These limitations motivate cautious interpretation and point to future work: longer and broader deployments, patient-level outcome studies, evaluation with multimodal CDS, and cost–benefit analyses that incorporate governance overheads and retraining pipelines.

9. Ethical Considerations

Deploying AI-mediated decision support in safety-critical care raises ethical questions beyond interface comfort or productivity. We organize considerations across *patient welfare*, *fairness and bias*, *autonomy and informed use*, *accountability and auditability*, *data governance and privacy*, and *sustainability and procurement*.

9.1. Patient Welfare and Nonmaleficence

Any interface that changes attention patterns can inadvertently worsen outcomes by masking rare but deadly conditions or by normalizing risky defaults. Our design avoids persuasive language that could overstate certainty, surfaces explicit uncertainty ranges, and preserves clinician override with minimal friction. We recommend unit-level safety gates for high-risk recommendations, prospective monitoring for near misses, and rapid rollback procedures.

9.2. Fairness and Bias

Risk models may encode historical inequities (e.g., triage thresholds influenced by differential access to care). Interfaces can amplify or mitigate bias depending on which evidence they foreground and how uncertainty is framed. We advocate subgroup-aware monitoring dashboards, explanation views that highlight sparse or low-quality inputs, and prompts that encourage clinicians to seek corroborating evidence for borderline cases. Governance should require bias audits before and after deployment, with clear escalation pathways when disparities are detected.

9.3. Autonomy, Informed Use, and Consent

Clinicians must understand the *intended use* and *limitations* of CDS. Progressive disclosure includes at-a-glance model scope, training data windows, and known exclusion criteria. For patients, transparency about AI involvement varies by jurisdiction and institutional policy; where feasible, consent materials should explain the role of decision support and data use in plain language.

9.4. Accountability and Auditability

When AI outputs are actionable, accountability becomes distributed among model developers, health IT teams, and clinicians. Our governance layer captures immutable logs linking inputs, outputs, user actions, and outcomes, supporting incident analysis without presupposing blame. Audit trails must be protected from tampering yet accessible to safety officers; retention schedules should align with medico-legal requirements.

9.5. Data Governance and Privacy

Evidence cards draw from multiple systems (EHR vitals, labs, notes). Aggregation can expand the attack surface. We recommend least-privilege data flows, encryption in transit and at rest, and minimization (showing only necessary elements). Feedback mechanisms must avoid inserting PHI into free-text fields where unnecessary; structured feedback (e.g., coded reasons for override) reduces privacy risk and improves learnability for model updates.

9.6. Sustainability and Responsible Procurement

Institutions should evaluate life-cycle costs and environmental impact (compute, storage, retraining). Preference should be given to models and interface layers that support efficient inference, scheduled retraining aligned to drift, and clear end-of-life plans. Procurement should require vendors to provide transparency artifacts (model cards, data sheets) and to commit to security updates.

9.7. Ethical Review and Community Participation

IRB review covered our studies; ongoing ethics should include frontline staff in design governance, patient representatives on oversight committees, and regular forums for surfacing harms and benefits. Interfaces must include easy channels to report safety concerns that route to accountable stewards.

9.8. Summary

Ethical deployment hinges on designing for uncertainty, preserving human agency, auditing for equity, and institutionalizing transparent accountability. Interface craft alone is insufficient without organizational commitments and resources.

10. Conclusion

This work frames clinical decision support as a human–computer interaction problem and proposes a three-layer design approach—information, coordination, and governance—that demonstrably improves calibrated trust and reduces unnecessary diagnostic revisions without inflating decision time. Across contextual inquiry, controlled comparisons, and field deployment, three patterns consistently helped: (1) coupling linguistic uncertainty labels with numeric ranges, (2) offering contrastive/counterfactual explanations aligned with clinical reasoning, and (3) progressively disclosing evidence to match evolving information needs.

Beyond immediate performance, the governance layer operationalizes transparency (provenance, audit trails) and post-deployment learning (lightweight feedback loops). These mechanisms are crucial for safe scaling across settings and for responsible response to drift and disparity.

10.1. Implications for Practice

Hospitals seeking to operationalize CDS should: (i) treat interface and governance requirements as first-class in procurement, (ii) budget for monitoring and retraining as continuing costs, (iii) pilot with mixed methods that combine usability metrics with safety reviews, and (iv) create escalation pathways for bias and safety issues.

10.2. Future Work

We aim to: (a) extend design patterns to multimodal CDS (imaging + text) and mobile contexts, (b) evaluate long-term patient-level outcomes, (c) study organizational adoption dynamics (training, incentives, safety culture), and (d) develop open reference implementations of provenance capture and explainability that can be audited and reused across institutions.

10.3. Closing Remark

As CDS diffuses into everyday care, success will depend less on isolated model scores and more on the sociotechnical craft of *making the right information show up at the right time, in the right form, with the right accountability.* Our results offer concrete steps toward that goal.

Acknowledgments

We thank the clinicians, nurses, and administrators who participated in observations, studies, and pilots; the health IT teams for integration support; and our institutional review boards for timely guidance. We also appreciate the anonymous reviewers whose feedback sharpened the framing and analysis. Any errors remain our own.

References

References

[1] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[3] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4765–4774, 2017.

[4] Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

[5] Shahin Tonekaboni, Shalmali Joshi, Melissa McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical decision support. *Machine Learning for Healthcare*, PMLR 106:359–380, 2019.

[6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD*, pages 1721–1730, 2015.

[7] Mark P. Sendak, Michael D'Arcy, Suresh Kashyap, et al. A path for translation of machine learning products into healthcare delivery. *JAMIA*, 27(12):1961–1967, 2020.

[8] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew Beam. The false hope of current approaches to explainable AI in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, et al. Beyond accuracy: The role of mental models in human–AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2019.

[10] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 126–137, 2015.

[11] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, et al. Manipulating and measuring model interpretability. *CHI '21*, Article 273, 2021.

[12] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[13] Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. Model cards for model reporting. In *Proceedings of the FAT* Conference*, pages 220–229, 2019.

[14] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[15] Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning. In *Machine Learning for Healthcare*, PMLR 106:547–558, 2019.

[16] Federico Cabitza, Andrea Rasoini, and Gianfranco Gensini. Unintended consequences of machine learning in medicine. *JAMA*, 318(6):517–518, 2017.

[17] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference*, pages 5092–5103, 2016.

[18] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng (Polo) Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE TVCG*, 25(8):2674–2693, 2019.

[19] Jenna Wiens and Mark S. D. Shenoy. Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, 66(1):149–153, 2018.

[20] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.