Trustworthy Multi-Modal AI in Healthcare: A Comprehensive Framework for Bias Detection, Explanation, and Mitigation

Author: Daniel Anderson, **Affiliation:** Lecturer, Faculty of Computer Science, Technical University of Munich, Germany. **Email:** daniel.anderson@tum.de

Abstract

Machine learning (ML) and multi-modal artificial intelligence (AI) promise transformative improvements in healthcare: earlier diagnosis, personalized treatment, and system-level efficiency. Yet these promises are coupled with well-documented risks algorithmic bias, opacity, and fragile generalization that can exacerbate health inequities and undermine trust. This paper proposes a principled, research-grade framework for *trustworthy multi-modal AI in healthcare* that integrates (1) systematic bias detection across modalities (imaging, structured EHR, clinical text, genomics), (2) causal and statistical explanation tools, and (3) layered mitigation strategies (preprocessing, in-training constraints, post-hoc adjustments, and socio-technical governance). We ground the framework in recent empirical failures and methodological advances, present an extensible software + evaluation pipeline, specify datasets and metrics for rigorous benchmarking (including subgroup calibration and counterfactual tests), and describe deployment-level governance aligned with WHO, EU and clinical reporting standards. Implementation recommendations emphasize reproducible code, federated/private training options, adversarial robustness checks, and continuous monitoring. Throughout we highlight tradeoffs (fairness definitions, utility vs. parity, explainability vs. fidelity) and provide concrete protocols researchers and health systems can adopt to reduce harms and make multi-modal AI clinically trustworthy.

Keywords: trustworthy AI, healthcare, bias detection, bias mitigation, multimodal machine learning, explainability, fairness, federated learning, algorithmic auditing

1. Introduction

Artificial intelligence that combines multiple data modalities medical imaging, structured electronic health records (EHRs), clinical notes, and molecular data holds exceptional promise for improving diagnostic accuracy and personalized care (Baltrusaitis, Ahuja, & Morency, 2018). Multimodal models can capture complementary signals that single-modality models miss, and their adoption in healthcare is accelerating across clinical domains (e.g., radiology, pathology, intensive care). However, empirical analyses reveal that deployed clinical algorithms can reproduce and amplify structural inequities; a notable example found that an algorithm used to

allocate extra care systematically disadvantaged Black patients by using healthcare costs as a proxy for medical need (Obermeyer et al., 2019). That study and others demonstrate how algorithmic design choices and data artifacts can produce clinically meaningful bias (Baltrušaitis, Ahuja, & Morency, 2017)

Trustworthy AI in healthcare therefore requires integrated technical and governance solutions: methods to detect and quantify bias across modalities, causal and local explanation techniques to attribute harms and spurious correlations, mitigation strategies that operate at data, model, and decision-making layers, and institutional processes for auditing, documentation, and post-deployment monitoring. This paper develops such a framework and supplies an evaluation and implementation roadmap tailored to the multi-modal clinical setting.

(We adopt the WHO and EU conceptions of "trustworthy AI," emphasizing human-centered design, technical robustness, transparency, privacy, fairness, and accountability; see Section 8.)

2. Background and problem framing

2.1 Why multimodal healthcare AI is different

Clinical decision-making is intrinsically multi-modal: a clinician combines lab results, patient history, imaging, and patient preferences. Multimodal ML aims to emulate this process by fusing heterogeneous inputs (Baltrusaitis et al., 2018). In healthcare, multimodal models can improve sensitivity and specificity, but they also increase surface area for bias: each modality can bring different collection practices, missingness patterns, and social confounders (e.g., differential imaging device usage across hospitals) (Fatunmbi, 2021). The literature on multimodal biomedical fusion highlights novel technical challenges alignment, co-training, missing modalities, and joint representation learning that interact with fairness and interpretability requirements (Kline et al., 2022; Baltrusaitis et al., 2018).

2.2 Evidence of harm and shortcut learning

Clinical AI failures fall into two broad patterns: (a) models learning *spurious shortcuts* (dataset artifacts, hospital identifiers, acquisition markers) that do not generalize (Geirhos et al., 2020; DeGrave et al., 2021), and (b) models that encode proxies for protected attributes (e.g., cost as proxy for need) which produce inequitable allocations (Obermeyer et al., 2019). These empirical patterns motivate a defensive design: rigorous external validation, careful dataset curation, and explicit bias audits using subgroup and counterfactual tests.

2.3 Fairness definitions and tradeoffs

Fairness is not a single metric competing definitions (demographic parity, equalized odds, equal opportunity, calibration) impose different constraints and tradeoffs with accuracy (Hardt, Price, & Srebro, 2016). In healthcare, where calibration and risk-rank ordering often matter (e.g., triage thresholds, treatment allocation), equalized odds or subgroup-conditional calibration may be more

clinically appropriate than demographic parity. Our framework treats fairness definitions as policy choices constrained by clinical context and legal regulation, not merely technical knobs (Fatunmbi, 2021).

3. Related technical literature

This section situates the framework in key methodological work:

- **Interpretability tools:** Local (SHAP, LIME), influence functions, and prototype/contradiction analyses help attribute model decisions to features or training examples (Lundberg & Lee, 2017; Koh & Liang, 2017). These are essential for clinician-facing explanations and for discovering spurious correlations.
- **Bias detection & toolkits:** Comprehensive toolkits and metrics (IBM AIF360, Microsoft Fairlearn) offer standardized fairness checks and mitigation primitives (Bellamy et al., 2018). These toolkits are helpful starting points but must be adapted to multimodal clinical data and to calibration-sensitive tasks.
- **Mitigation algorithms:** Preprocessing (reweighing, massaging), in-training constraints (adversarial debiasing, constrained optimization), and post-hoc adjustments (thresholding, calibrated equalized odds) are all viable, with different cost/benefit profiles (Kamiran & Calders, 2012; Zhang, Lemoine, & Mitchell, 2018).
- **Federated and privacy-preserving training:** Federated learning and differential privacy offer pathways for training robust models on distributed clinical data without sharing patient records, which also helps generalizability and may reduce dataset shift but introduces new challenges for fairness evaluation and debugging (Sheller et al., 2020; Li et al., 2019).
- **Multimodal fusion & robustness:** Surveys and empirical work show both promise and pitfalls for multimodal fusion; understanding modality-specific bias and interactions between modalities is critical for trustworthy systems (Baltrusaitis et al., 2018; Warner et al., 2024).

4. A Multi-Modal Trustworthy AI Framework (MM-Trust)

We propose *MM-Trust*, a structured, modular framework that combines technical components, evaluation protocols, and governance practices. MM-Trust is organized in three interacting layers:

1. **Data & Preprocessing layer** modality-aware bias scanning, harmonization, synthetic augmentation, and privacy protections.
2. **Modeling & Explanation layer** multimodal fusion architectures with built-in fairness regularizers, explainability tools (global/local), influence-based debugging, and adversarial robustness checks.

3. **Deployment & Governance layer** documentation (model cards, data sheets), continuous monitoring, clinical validation protocols (TRIPOD+AI / CONSORT-AI), and stakeholder governance.

Below we detail each layer and provide concrete algorithms and protocols.

4.1 Data & Preprocessing layer

4.1.1 Modality-aware bias scanning.

For each modality (imaging, structured EHR, text, genomics), run a battery of audits:

- *Representation audit:* compute subgroup prevalences across protected attributes (race, gender, age, socioeconomic proxies) and across sites; flag under-represented groups (Fatunmbi, 2022).
- *Missingness audit:* analyze missingness mechanisms (MCAR, MAR, MNAR) by subgroup and modality; many clinical biases appear as non-random missingness.
- *Acquisition artifact audit (imaging):* detect explicit acquisition or site markers (watermarks, machine headers) using shallow classifiers trained to predict site; high site predictability signals potential shortcut risks (Samuel, 2021).
- *Proxy audit:* identify plausible proxies (cost, prior utilization) for protected traits and for the target. If a commonly used feature is a proxy for disadvantage (e.g., cost as a proxy for health need), the feature should be carefully re-evaluated or replaced.

These audits combine statistical tests, small diagnostic classifiers, and clinician review. Tools: AIF360, Fairlearn, custom audit scripts.

4.1.2 Harmonization & augmentation.

Standardize coding systems (ICD/LOINC), map vocabularies across sites, and apply realistic augmentation (image contrast/hardware simulation; synthetic EHR generation) to mitigate dataset shift. For sensitive groups that are under-represented, oversampling should be used cautiously (avoid synthetic label noise); where feasible, prioritize acquiring additional real data (Samuel, 2022).

4.1.3 Privacy & provenance.

Adopt provenance metadata (who collected, when, device, preprocessing steps) and de-identification standards. Where centralization is infeasible, adopt federated learning or secure aggregation; log all provenance and compute metadata to enable audits.

4.2 Modeling & Explanation layer

4.2.1 Multimodal fusion architectures.

Choose fusion strategy based on missingness and alignment:

- *Early fusion* (concatenate feature representations) works when modalities are consistently present and aligned.
- *Late fusion / ensemble* is preferred when modalities are intermittently missing or when interpretability of each modality is critical.
- *Hybrid attention-based fusion* (modality-specific encoders + cross-modal transformers or attention modules) is powerful for learning interactions but requires larger data and careful monitoring for shortcut reliance.

For each architecture, incorporate *modality-specific fairness constraints*. Example: add adversarial branch that attempts to predict protected attributes from the fused representation; minimize predictive objective while minimizing protected-attribute predictability (adversarial debiasing). Zhang et al. (2018) provide an implementation template.

4.2.2 Explainability: global and local

Adopt a layered explanation strategy:

- *Global explanations* (feature-importance aggregated across populations using SHAP, concept activation vectors) reveal overall model reliance patterns (Lundberg & Lee, 2017).
- *Local explanations* (SHAP, influence functions) are necessary for per-patient justification and for debugging individual errors (Koh & Liang, 2017).
- *Counterfactual explanations* (minimal change to input that flips prediction) are particularly valuable in clinical settings to suggest actionable interventions.

4.2.3 Robustness and shortcut detection

Implement active tests for shortcuts:

- *Domain classification probe:* train a classifier on representation to predict site/device/source; high accuracy indicates leakage of non-clinical signals.
- *Out-of-distribution (OOD) stress tests:* measure performance on holdout hospitals, demographic subgroups, and simulated confounding variants.
- *Ablation/feature-removal tests:* measure performance drop when removing suspect features (e.g., cost). These tests helped reveal the Obermeyer cost-proxy failure.

4.3 Deployment & Governance layer

4.3.1 Documentation and reporting

Produce machine-readable and clinician-readable artifacts: model cards, data sheets for datasets, and audit logs. Adhere to clinical AI reporting standards (SPIRIT-AI / CONSORT-AI for trials; TRIPOD+AI for prediction model reporting) to ensure transparency and reproducibility.

4.3.2 Monitoring and feedback loops

Deploy continuous monitoring for distribution shifts (covariate drift, label shift), subgroup performance degradation, and real-world harms. Maintain a rapid-response governance process (data safety monitoring board for AI) for remediation (Samuel, 2021).

4.3.3 Governance committees and clinician oversight

Establish multidisciplinary oversight with clinicians, ethicists, statisticians, and patient representatives. Define acceptable thresholds for subgroup calibration errors, unacceptable differences in false negative rates, and escalation paths.

5. Evaluation protocol and metrics

A rigorous evaluation plan is crucial. Below we specify a hierarchy of metrics and experimental design tailored to multimodal clinical tasks.

5.1 Experimental design

- **Train / validation / external test split by site**: ensure splits respect institution to detect generalization failures.
- **Subgroup holdouts**: evaluate by protected groups (e.g., race, sex, age) and by intersectional groups (e.g., older Black women).
- **Counterfactual test sets**: where feasible, construct or simulate counterfactual examples to probe causal sensitivity.
- **Stress testing**: test on corrupted/noised variants (imaging artifacts, missing labs) to assess robustness.

5.2 Core metrics

- **Predictive performance:** AUROC, AUPRC, MAE/RMSE for regression; but always report **calibration** (Brier score, calibration plots), because decision thresholds in healthcare depend on well-calibrated probabilities.
- **Fairness metrics:** equalized odds difference, equal opportunity gap (Hardt et al., 2016), demographic parity where relevant, subgroup calibration error, and precision/recall by subgroup.
- **Explainability diagnostics:** stability of feature importances (SHAP), influence-function outliers, and explanation-to-outcome concordance.
- **Operational metrics:** time-to-prediction, resource consumption, and human-in-the-loop latency.
- **Economic / clinical utility:** expected misclassification cost (treatment harms), decision curve analysis, number needed to treat / harm.

5.3 Audit checklist (minimum)

1. Data provenance log and distribution across protected groups.
2. Site-by-site performance and calibration plots.
3. Feature-importance and shortcut detection report (site classifier accuracy).
4. Fairness metric table with mitigation applied and pre/post comparisons.
5. Documentation: model card + decision thresholds and clinical integration notes.

6. Bias detection toolbox: methods and recipes

Below are practical, reproducible recipes for commonly encountered bias patterns.

6.1 Detecting proxy-based allocation bias

Problem: model uses a proxy (e.g., cost) that is correlated with the target but biased by access differences.
Recipe:

1. Compute correlation of suspect feature with protected attributes and with outcomes.
2. Build a *proxy test*: train a model excluding the suspect feature; if subgroup performance improves or parity increases, the suspect feature likely drives bias.
3. Counterfactual simulation: replace suspect features with values drawn from advantaged subgroup distribution and observe outcome shifts.

(Obermeyer et al. 2019 used this reasoning to show cost was a poor proxy for need.)

6.2 Shortcut / confounder discovery (imaging + EHR)

Problem: model learns imaging or site artifacts.
Recipe:

1. Train a *site classifier* on learned representations; high accuracy → shortcut risk.
2. Use explainability (saliency, integrated gradients, SHAP) to find image regions most responsible. If non-anatomical regions dominate, perform data cleaning and retrain.
3. Perform external validation across hospitals with different acquisition devices.

(DeGrave et al. 2021 showed that COVID chest X-ray models learned dataset origin shortcuts rather than pathology.)

6.3 Subgroup calibration and fairness tests

Recipe:

1. Compute calibration curves and Brier scores per subgroup.
2. Compute equalized odds differences (FPR/FNR gap) and equal opportunity gaps.
3. When misalignment exists, prioritize calibration (recalibration by subgroup) when treatment decisions depend on probability thresholds.

7. Bias mitigation strategies: mapping to clinical use cases

Mitigation must be chosen with awareness of clinical tradeoffs. Below is a decision matrix and practical guidance.

7.1 Preprocessing (data level)

- **When to use:** when bias originates from sampling or label artifacts (e.g., under-representation).
- **Methods:** reweighing, targeted oversampling, label massaging for historical bias correction (Kamiran & Calders, 2012).
- **Pros/cons:** conceptually simple; may not fix causal confounding; synthetic oversampling risks creating unrealistic clinical combinations.

7.2 In-training constraints

- **Adversarial debiasing:** minimize prediction loss while an adversary attempts to recover protected attributes from representations. Works well for many tasks but can reduce performance and be unstable; needs hyperparameter tuning and clinical validation (Zhang et al., 2018).
- **Constrained optimization:** impose equality constraints (e.g., equalized odds) directly during training (Hardt et al., 2016).

7.3 Post-processing

- **Thresholding and calibration:** adjust decision thresholds per subgroup to equalize certain rates; suitable when probabilities are well-calibrated and legal/social context permits group-aware thresholds.
- **Reject option classification:** defer uncertain cases to human review, useful in high-risk contexts.

7.4 Socio-technical mitigation

- **Human-in-the-loop triage:** preserve clinician oversight and require explicit acceptance of model recommendations in sensitive scenarios.

- **Policy change:** where models reflect structural inequity (e.g., access gaps), remediation may require non-technical interventions (resource allocation changes).

8. Governance, standards, and clinical trial reporting

Regulatory and standards frameworks increasingly require transparency, human oversight, and reproducibility. WHO and EU guidelines frame trustworthy AI around human-centeredness, technical robustness, privacy, transparency, fairness, and accountability (WHO, 2021; EU HLEG, 2019). Clinical reporting extensions (SPIRIT-AI / CONSORT-AI; TRIPOD+AI) now specify reporting items for AI interventions and prediction models; compliance aids peer review and adoption. Model documentation (model cards, data sheets) should be standard outputs for any clinical AI deployment.

9. Implementation considerations and a recommended software stack

Software tools & libraries

- Multimodal training: PyTorch / TensorFlow; Hugging Face Transformers for text encoding; MONAI for medical imaging.
- Fairness & audit tools: IBM AIF360, Microsoft Fairlearn, SHAP for explanations.
- Privacy/federation: Flower, TensorFlow Federated, PySyft for federated training; Opacus for differential privacy.
- Reproducibility: containerized experiments (Docker), versioned datasets (DVC), seed logging, and provenance metadata.

Compute and operational notes

- Multimodal transformers and attention models can require large compute and careful regularization; when compute or dataset size is limited, prefer modular late-fusion architectures.
- Federated training reduces raw data sharing but complicates debugging and fairness audits design federated evaluation protocols in advance.

10. Case study (illustrative): chest X-ray + EHR triage model

A hospital wishes to deploy a multimodal triage model combining CXR images and recent EHR features to prioritize chest pain admissions. Following MM-Trust:

1. **Data audit:** site distribution shows devices A/B/C; age and race distributions differ by site (representation audit).
2. **Shortcut probe:** site classifier on representations yields 0.92 accuracy → clear site leakage.

3. **Mitigation:** retrain using domain-invariant representation learning (adversarial debiasing + data harmonization), then apply subgroup calibration and deploy a human-in-loop reject policy for high-uncertainty cases.
4. **Evaluation:** external validation on held-out hospital yields reduced AUROC drop and improved subgroup calibration; continuous monitoring is instituted.

This case illustrates the end-to-end application of detection, mitigation, and governance.

11. Limitations, open problems, and research directions

- **Causal inference in multimodal data**: bias that arises from causal pathways (e.g., differential treatment causing outcome differences) requires causal models and often external interventions; purely statistical mitigation can be insufficient.
- **Federated fairness auditing**: auditing fairness under privacy constraints is technically challenging and remains an active area.
- **Explainability guarantees**: contrast between human-useful explanations and faithful explanations needs more empirical work in clinical settings.
- **Benchmarking and incentives**: the field needs standardized multimodal datasets with rich demographic metadata and well-defined ground truth to benchmark fairness interventions.

These are fertile areas for further work (Mehrabi et al., 2021; Warner et al., 2024).

12. Practical checklist for practitioners (summary)

1. **Before modeling:** provenance logs, representation & missingness audits, clinical stakeholder review.
2. **During modeling:** site probes, adversarial or constrained fairness techniques as appropriate, robust fusion choices.
3. **Pre-deployment:** external validation by site and subgroup, TRIPOD+AI / CONSORT-AI checklist, model card & clinician training.
4. **Post-deployment:** continuous monitoring, reporting, rapid remediation workflow, and periodic re-audits.

13. Conclusion

Multi-modal AI offers powerful opportunities for healthcare but creates commensurate responsibilities. Trustworthy deployment requires an integrated, multi-layered approach combining careful data curation, modality-aware modeling, explainability and targeted mitigation, and strong governance aligned with clinical and regulatory norms. The MM-Trust framework offered here is pragmatic: it prescribes concrete detection tests, modeling practices, evaluation

protocols, and governance steps that teams can adopt today to reduce harms while enabling innovation. With rigorous experimentation, standardized benchmarks and cross-disciplinary oversight, multimodal AI can be made both powerful and trustworthy in clinical practice.

References

1. Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias* (arXiv:1810.01943). arXiv.
2. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.
3. DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3, 610–619.
4. Fatunmbi, T. O. (2021). Integrating AI, Machine Learning, and Quantum Computing for Advanced Diagnostic and Therapeutic Strategies in Modern Healthcare. *International Journal of Engineering and Technology Research*, 6(1), 26–41. https://doi.org/10.34218/IJETR_06_01_002
5. Fatunmbi, T. O. (2022). Impact of data science and cybersecurity in e-commerce using machine learning techniques. *World Journal of Advanced Research and Reviews*, 13(1), 832–846. https://doi.org/10.30574/wjarr.2022.13.1.0607
6. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 665–673.
7. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (NeurIPS).
8. Kamiran, F., & Calders, T. G. K. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8.
9. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning* (ICML).
10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions (SHAP). In *Advances in Neural Information Processing Systems* (NeurIPS).

11. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6, 26094.

12. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342.

13. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *The New England Journal of Medicine*, 380, 1347–1358. *(review)*

14. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215.

15. Samuel, A. J. (2021). Cloud-Native AI solutions for predictive maintenance in the energy sector: A security perspective. *World Journal of Advanced Research and Reviews*, 9(03), 409–428. https://doi.org/10.30574/wjarr.2021.9.3.0052

16. Samuel, A. J. (2022). AI and machine learning for secure data exchange in decentralized energy markets on the cloud. *World Journal of Advanced Research and Reviews*, 16(2), 1269–1287. https://doi.org/10.30574/wjarr.2022.16.2.1282

17. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., Bakas, S. (2020). Federated learning in medical imaging: A systematic review. *Scientific Reports*, 10, 1179.

18. Warner, E., Lee, J., Hsu, W., Syeda-Mahmood, T., Kahn, C., Gevaert, O., & Rao, A. (2024). Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects. *npj Digital Medicine*.

19. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.