When LLMs Decide Who Gets Care: A Vision for Multi-Agent Systems in High Stakes Clinical Decision-Making

Divya Sharma
Department of Mathematics and Statistics, York University
Toronto, Ontario, Canada
divya03@yorku.ca

Ghazal Azarfar*
Toronto General Hospital, University Health Network
Toronto, Ontario, Canada
ghazal.azarfar@uhn.ca

Bima Hasjim*
Department of Surgery, University of California – Irvine
Orange, California, USA
bhasjim@hs.uci.edu

Mamatha Bhat
Toronto General Hospital, University Health Network
Toronto, Ontario, Canada
mamatha.bhat@uhn.ca

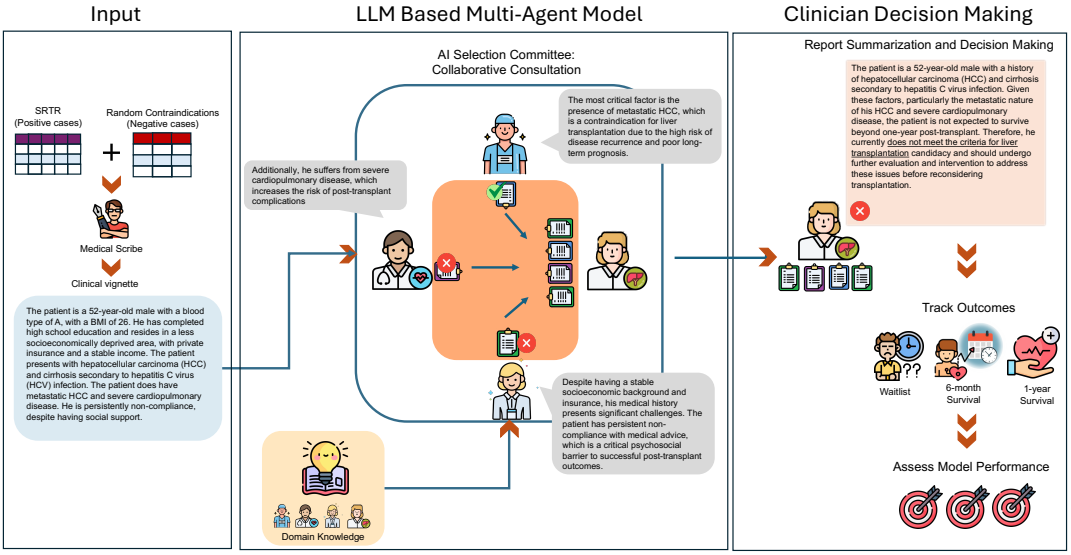

Figure 1: Overall framework for the multi-agent modeling approach designed to support clinical decision-making.

Abstract

As large language models (LLMs) transition from static predictors to autonomous agents, a promising application lies in simulating real-world, multi-disciplinary clinical committees responsible for life-or-death decisions such as organ transplant eligibility. This vision paper explores the design and deployment of multi-agent LLM systems that emulate role-specialized clinicians collaborating to assess high-stakes medical cases. Using empirical insights from a simulation of a liver transplant selection committee, we highlight how even highly accurate agents can propagate disparities in the absence of subgroup-sensitive reasoning and explainability. We argue that multi-agent LLMs must go beyond role emulation to incorporate counterfactual rationalization, inter-agent transparency, and clinician-in-the-loop arbitration. Our vision sets forth a roadmap for building accountable, equitable, and trustworthy multi-agent LLM systems that can support, not replace critical clinical deliberation.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Multi-Agent Systems, Clinical Decision-Making, Liver Transplant, Fairness in AI, Explainability, Large Language Models (LLMs)

ACM Reference Format:
Divya Sharma, Ghazal Azarfar, Bima Hasjim, and Mamatha Bhat. 2025. When LLMs Decide Who Gets Care: A Vision for Multi-Agent Systems in High Stakes Clinical Decision-Making. In . ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

As large language models (LLMs) evolve into autonomous agents capable of reasoning, memory, and collaboration, their use in high-stakes domains like medicine is accelerating [1, 2]. One frontier application is the simulation of clinical committees through multi-agent LLM systems collections of specialized agents that mirror real-world medical roles such as transplant surgeons, hepatologists, or social workers [3, 4]. These systems promise to standardize clinical decision-making, expand access to expert-level evaluation, and support scalability in resource-constrained settings [5, 6].

However, with this promise comes risk. In medical contexts, especially in life-critical domains like organ transplantation, decisions made by clinicians determine who receives limited life-saving treatments. Transplant eligibility decisions, for example, require nuanced judgment that balances clinical prognosis with psychosocial and ethical considerations [7, 8, 9]. In the United States, these decisions are made by multidisciplinary transplant committees. Simulating such committees with LLM agents introduces a new sociotechnical paradigm where AI not only assists, but potentially decides who gets care.

This paper presents a vision for fair and explainable multi-agent LLM systems in clinical settings, grounded in a case study of liver transplant eligibility assessment. We simulate over 8,000 patient evaluations using a committee of LLM-based agents each role-tuned and interacting under orchestrated clinical protocols. Despite achieving high accuracy in identifying medical contraindications and predicting survival, the system exhibited systematic disparities. Patients from marginalized groups particularly women, Black individuals, and those with low socioeconomic status received disproportionately lower eligibility scores. These disparities stemmed from agents' reliance on social proxy features such as insurance status, education level, and the Area Deprivation Index (ADI), compounded by the absence of patient-specific rationale or counterfactual reasoning [7, 8, 9].

Moreover, the architecture of multi-agent LLM systems introduces novel failure modes. Each agent operates semi-autonomously, producing decisions without full visibility into the others' reasoning. While this mirrors real clinical practice, it creates challenges for interpretability and post hoc auditing critical components for trust in clinical AI [10, 11]. Unlike single-model classifiers, these systems distribute judgment across agents with disjoint observations and distinct objectives, making system-level transparency difficult to achieve [12, 13].

This vision paper identifies these architectural limitations and proposes a new agenda for agentic AI in healthcare one that embeds fairness constraints, supports case-level explainability, and enables clinician-in-the-loop oversight. Our broader contribution is to argue that in domains where AI systems influence or decide who receives care, design choices must be driven by equity and accountability from the outset [14, 15]. Multi-agent AI systems, if left unchecked, risk encoding and legitimizing historical disparities under the veneer of objectivity. We propose a design framework that combines subgroup-aware model training, counterfactual simulation modules, and explainability stacks with deployment protocols that mitigate the digital divide. In the sections that follow, we summarize our empirical findings from the transplant simulation, review key failure modes in multi-agent clinical AI, and propose a roadmap for building systems that are not only high-performing, but also just and transparent.

2 Empirical Insight: A Multi-Agent Simulation of Transplant Committee Decision-Making

To understand the practical implications of agentic AI in high-stakes healthcare settings, we designed a simulated transplant selection committee using multiple role-specialized LLM agents tasked with evaluating and selecting patients for transplantation. This section provides the empirical grounding for our proposed vision, drawing on a large-scale, retrospective cohort of liver transplant recipients in the U.S. and structured simulation of committee-based decision-making.

2.1 Data Source and Cohort

We leveraged data from the Scientific Registry of Transplant Recipients (SRTR), a national U.S. registry encompassing all candidates and recipients of solid organ transplants. The study cohort included adult patients (≥18 years) who underwent deceased donor liver transplantation (DDLT) between January 2004 and June 2024.

To test the robustness of the AI system under complex clinical conditions, 16.4% of the cohort was augmented with synthetic contraindication profiles randomly assigned from a set of medically recognized ineligibility criteria, including metastatic hepatocellular carcinoma (HCC), extrahepatic malignancies, severe cardiopulmonary comorbidities, active substance use, and persistent noncompliance.

Each case comprised 59 variables, including standard clinical indicators (e.g., MELD score, liver disease etiology), demographic factors (age, sex, race/ethnicity), and social determinants of health (insurance type, education level). We additionally incorporated Area Deprivation Index (ADI), a validated geospatial measure of neighborhood-level socioeconomic status, which was merged via ZIP code using publicly available census tract data.

2.2 Multi-Agent Architecture and Role Configuration

To simulate multidisciplinary clinical deliberation, we developed a five-agent architecture using GPT-4 models orchestrated through CrewAI (v0.63.6) and LangChain. Agents were assigned the following roles: transplant hepatologist (committee chair), transplant surgeon, cardiologist, social worker, and medical scribe. Each domain-specialized agent received natural language vignettes created by the scribe, which translated structured SRTR data into narrative form.

Prompts were carefully engineered for domain fidelity and included zero-shot and chain-of-thought reasoning variants. A self-consistency protocol sampling each agent's decision five times and returning the modal output was applied to enhance decision stability. Committee voting was based on majority consensus, with the hepatologist agent responsible for adjudicating ties, reflecting clinical governance norms in transplant boards.

2.3 Performance and Fairness Assessment

The AI-staffed committee demonstrated high accuracy in core clinical tasks. Across 8,412 patient vignettes, the system achieved 98.2% accuracy in detecting transplant contraindications and over 94% accuracy in predicting post-transplant survival at both 6 and 12 months.

However, subgroup analysis revealed concerning patterns of inequity:

- **Disparate Impact (DI):** Female patients had a DI of 0.78 compared to male patients; Black patients 0.85 compared to White; patients with high school education or less scored 0.82; and those from the most socioeconomically deprived areas (ADI quintile 5) had a DI of 0.64. All scores fell below the accepted fairness threshold of 0.80, suggesting systematic under-selection.
- **Attribution Patterns:** Cosine similarity analysis of agent rationales revealed domain-aligned behavior (e.g., cardiologists prioritized cardiovascular risk). However, the social worker agent placed disproportionate weight on non-clinical features particularly insurance type, education level, and ADI which are tightly coupled with structural inequities in access to care.

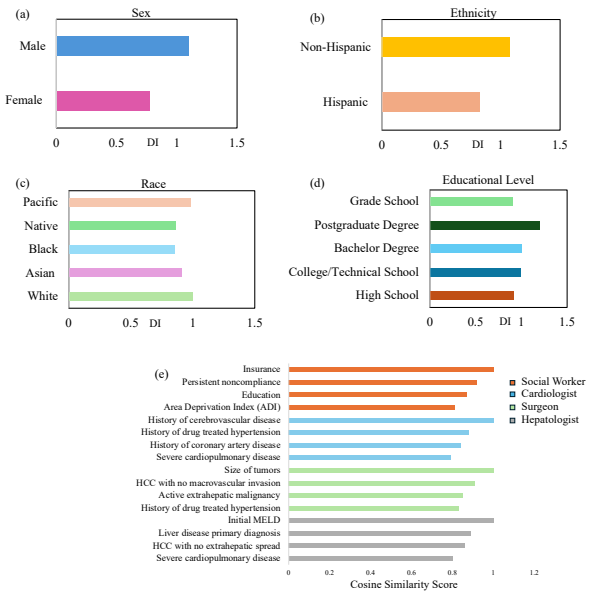

Figure 2: Disparity and Attribution Metrics in Multi-Agent AI: (a–d) Disparate Impact (DI) scores across sex, ethnicity, race, and education reveal consistent under-classifzication of disadvantaged groups. (e) Cosine similarity analysis highlights differential reliance on clinical vs. socioeconomic variables across agent roles.

2.4 Implications for Systemic Bias in Multi-Agent AI

While overall performance was strong, these findings illustrate a broader concern: agentic AI systems especially those structured to simulate real-world decision roles may inadvertently amplify institutional biases embedded in historical data. Without mechanisms for fairness optimization, transparency, and clinical oversight, these systems risk reinforcing rather than correcting disparities in care allocation.

Figure 2 illustrates both the disparity metrics (DI) across subgroups and role-specific attribution weights for selected features, highlighting how system-level disparities may emerge from the interaction of well-calibrated but siloed agents.

This simulation provides a critical empirical foundation for rethinking how agentic systems are designed, governed, and deployed in healthcare particularly in contexts where the stakes are nothing less than life or death.

2.5 Failure Modes

Despite strong diagnostic performance, our empirical evaluation uncovered three interrelated failure modes that limit the clinical viability, fairness, and transparency of multi-agent LLM systems in high-stakes healthcare applications.

(1) **Proxy Bias in Role-Specialized Reasoning.** Although each agent was prompted to emulate domain-specific expertise (e.g., cardiologist focusing on cardiac risk), feature attribution analyses revealed heavy reliance on non-clinical social proxies such as insurance status, education level, and Area Deprivation Index (ADI). These proxies, while correlated with clinical outcomes in the training data, are often entangled with historically marginalized group characteristics (e.g., race, gender, or socioeconomic status). For example, the "social worker" agent disproportionately penalized patients with public insurance or lower education, leading to systematically lower eligibility classifications for these groups. This reveals how role-aligned agents can amplify structural inequities under the guise of professional specialization.

(2) **Lack of Case-Level Explainability and Counterfactual Reasoning.** Although agents produced rationales in natural language, they lacked formalized attribution mechanisms (e.g., saliency maps, token-level importance) and could not support counterfactual queries such as: *"Would this patient have been accepted if their ZIP code indicated lower socioeconomic deprivation?"* This absence of trajectory-aware or causal reasoning means that even correct classifications lack auditable justification. In clinical domains where decisions often hinge on small differences in risk-benefit interpretation, this black-box opacity erodes trust, accountability, and recourse, especially for borderline cases.

(3) **Opaque Consensus Mechanisms and Intra-Agent Dependencies.** Final committee-level decisions were reached via majority vote, with the transplant hepatologist serving as a tie-breaker. However, there was no traceability into how agents influenced each other, nor visibility into inter-agent disagreement. This structure obscures whether disagreements stemmed from interpretive variance, uncertainty, or bias. In practice, such opacity undermines not only clinical accountability but also violates emerging AI governance requirements that demand decision traceability and dispute resolution mechanisms for AI-assisted diagnostics.

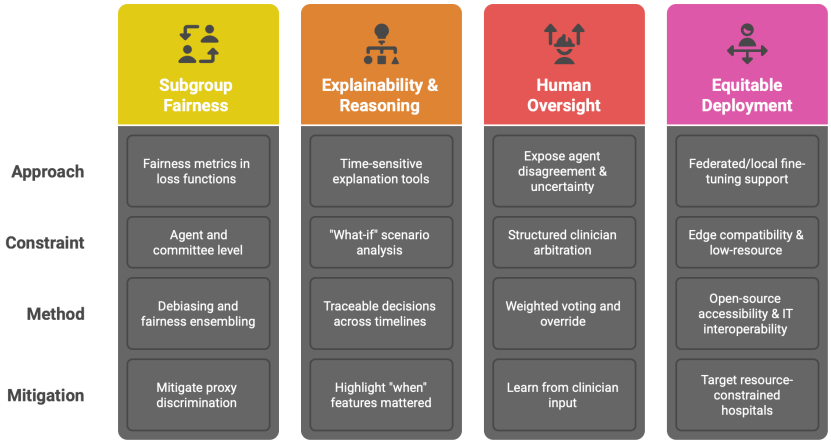

Figure 3: A modular blueprint for building equitable multi-agent LLM systems in medicine. Each pillar outlines key design considerations from fairness optimization and temporal explainability to clinician-in-the-loop oversight and infrastructure for equitable deployment necessary for responsible integration of Agentic AI into high-stakes clinical workflows.

Together, these failure modes illustrate that even high-performing agentic systems when deployed without explicit fairness constraints, transparency scaffolding, or human oversight can codify institutional biases, mask error pathways, and exclude vulnerable populations from life-saving interventions.

3 Vision: A Blueprint for Equitable Multi-Agent Systems

Our empirical findings highlight that achieving technical accuracy is not sufficient in high-stakes clinical settings. Multi-agent systems must be built with equity, transparency, and context-specific governance from the ground up. Below, we outline a blueprint with four interconnected pillars for developing responsible and socially aligned multi-agent LLM systems in medicine:

- **Subgroup-Aware Constraints and Fairness Optimization.** Traditional model evaluation metrics (e.g., accuracy, AUROC) mask disparities in treatment across protected subgroups. We advocate embedding fairness objectives such as Disparate Impact (DI), equalized odds, or subgroup calibration directly into agent loss functions [16]. In multi-agent settings, these constraints should operate both at the individual agent level and at the level of final consensus decisions. Approaches such as adversarial debiasing [17], representation-level regularization [18], and fairness-aware ensembling can help mitigate proxy discrimination (e.g., when education level/ insurance status stand in for socioeconomic status).
- **Trajectory-Aware Explainability and Counterfactual Modules.** Clinical decisions are temporally grounded and often hinge on trends (e.g., improving liver function) rather than single-timepoint data. Therefore, explanation modules must move beyond static feature attribution. We propose the integration of time-aware attribution methods (e.g., Temporal SHAP, Integrated Gradients) [19, 20] to highlight not just what features mattered, but when. Additionally, embedding conditional generative models such as CausalGANs [21] or counterfactual transformers [22] enables systems to answer "what-if" queries: Would a patient's eligibility status change if their insurance type or ZIP code were different?
- **Clinician-in-the-Loop Governance and Oversight.** In settings where AI influences human judgment, the burden of justification must be high. Multi-agent systems should expose internal disagreement, provide uncertainty estimates per agent, and support structured clinician arbitration. This may include interactive dashboards for reviewing agent rationales, weighted voting schemes based on domain confidence, and mechanisms to record and learn from clinician overrides [23]. Real-time flagging of low-confidence cases can help ensure human accountability in ambiguous scenarios.
- **Equitable Infrastructure and Deployment Protocols.** The benefits of agentic AI should not be limited to tertiary care centers or well-resourced institutions. Federated learning, domain adaptation, and privacy-preserving fine-tuning can enable site-specific calibration without centralizing protected health data [24]. We also emphasize importance of interoperability with existing systems, edge-compatible deployment for bandwidth-constrained hospitals, and open-source toolkits to reduce barriers to adoption. Equity must be embedded not only in model design, but in the ecosystem of tools, funding, and governance frameworks that enable sustained and inclusive use.

As large language models evolve from assistants to autonomous decision-makers, clinical AI governance must evolve as well. Fairness and explainability can no longer be treated as post hoc audits they must be engineered as first-class constraints in the design, training, and deployment of multi-agent healthcare systems. This blueprint offers a path forward: one that ensures such systems do not just mimic clinical reasoning, but elevate it in ways that are just, transparent, and aligned with institutional and societal values.

Code Availability: https://github.com/gazarfar/Liver-Transplant-AI-Agent-Committee/

References

[1] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29, 8, 1930–1940.

[2] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: how far are we from baymax? *arXiv preprint arXiv:2502.11211*.

[3] Muhammad Usman Tariq. 2024. Multi-agent models in healthcare system design. In *Bioethics of Cognitive Ergonomics and Digital Transition*. IGI Global, 143–170.

[4] Kai Chen, Xinfeng Li, Tianpei Yang, Hewei Wang, Wei Dong, and Yang Gao. 2025. Mdteamgpt: a self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856*.

[5] Sambasiva Rao Suura. 2025. Agentic ai systems in organ health management: early detection of rejection in transplant patients. *Journal of Neonatal Surgery*, 14, 4s.

[6] Mehrdad Rahsepar Meadi, Tomas Sillekens, Suzanne Metselaar, Anton van Balkom, Justin Bernstein, and Neeltje Batelaan. 2025. Exploring the ethical challenges of conversational ai in mental health care: scoping review. *JMIR mental health*, 12, e60432.

[7] Amit K Mathur, Douglas E Schaubel, Qi Gong, Mary K Guidinger, and Robert M Merion. 2010. Racial and ethnic disparities in access to liver transplantation. *Liver Transplantation*, 16, 9, 1033–1040.

[8] Amit K Mathur, Douglas E Schaubel, Qi Gong, Mary K Guidinger, and Robert M Merion. 2011. Sex-based disparities in liver transplant rates in the united states. *American Journal of Transplantation*, 11, 7, 1435–1443.

[9] Jaime M Glorioso. 2021. Kidney allocation policy: past, present, and future. *Advances in Chronic Kidney Disease*, 28, 6, 511–516.

[10] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366, 6464, 447–453.

[11] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2020. Chexclusion: fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*. World Scientific, 232–243.

[12] Soraia F Paulo, Isabel Neto, Nuno Leitão Figueiredo, Daniel Simões Lopes, and Hugo Nicolau. 2025. Reimagining multidisciplinary teams: challenges and opportunities for llms in cancer mdts. *Proceedings of the ACM on Human-Computer Interaction*, 9, 2, 1–22.

[13] Qingyun Wu et al. 2023. Autogen: enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

[14] Thomas Grote and Geoff Keeling. 2022. Enabling fairness in healthcare through machine learning. *Ethics and Information Technology*, 24, 3, 39.

[15] Pranav Rajpurkar and Eric J Topol. 2025. A clinical certification pathway for generalist medical ai systems. *The Lancet*, 405, 10472, 20.

[16] Lisa Koutsoviti Koumeri, Magali Legast, Yasaman Yousefi, Koen Vanhoof, Axel Legay, and Christoph Schommer. 2023. Compatibility of fairness metrics with eu non-discrimination laws: demographic parity & conditional demographic disparity. *arXiv preprint arXiv:2306.08394*.

[17] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

[18] Hao Yang, Zhining Liu, Zeyu Zhang, Chenyi Zhuang, and Xu Chen. 2023. Towards robust fairness-aware recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 211–222.

[19] Scott M Lundberg et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2, 10, 749–760.

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[21] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. 2017. Causalgan: learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*.

[22] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. 2022. Causal transformer for estimating counterfactual outcomes. In *International conference on machine learning*. PMLR, 15293–15329.

[23] Andrej Thurzo. 2025. Provable ai ethics and explainability in medical and educational ai agents: trustworthy ethical firewall. *Electronics*, 14, 7, 1294.

[24] Mansoor Ali, Faisal Naeem, Muhammad Tariq, and Georges Kaddoum. 2022. Federated learning for privacy preservation in smart healthcare systems: a comprehensive survey. *IEEE journal of biomedical and health informatics*, 27, 2, 778–789.