

# Chandra Wicaksono

## Melampaui Ekspektasi: Evaluasi Kinerja Dunia Nyata, Bias Algoritmik, dan Integrasi Kecerdasan Buatan dalam Triase Kli...

 Quick Submit Quick Submit Universitas 17 Agustus 1945 Semarang

### Document Details

Submission ID

trn:oid:::1:3645234035

Submission Date

Sep 10, 2026, 8:18 PM GMT+7

Download Date

Sep 10, 2026, 8:32 PM GMT+7

File Name

Chandra\_Wicaksono\_-\_jurnal.docx

File Size

118.4 KB

11 Pages

3,208 Words

22,253 Characters

# 22% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

## Top Sources

- 22%  Internet sources
- 0%  Publications
- 0%  Submitted works (Student Papers)

## Integrity Flags

### 0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

## Top Sources

22% Internet sources  
0% Publications  
0% Submitted works (Student Papers)

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

|    |          |                          |     |
|----|----------|--------------------------|-----|
| 1  | Internet | jurnal.muaraedukasi.id   | 8%  |
| 2  | Internet | formative.jmir.org       | 1%  |
| 3  | Internet | mjiri.iums.ac.ir         | 1%  |
| 4  | Internet | ouci.dntb.gov.ua         | <1% |
| 5  | Internet | pubmed.ncbi.nlm.nih.gov  | <1% |
| 6  | Internet | www.mashupmd.com         | <1% |
| 7  | Internet | arxiv.org                | <1% |
| 8  | Internet | repositorioceme.ufrgs.br | <1% |
| 9  | Internet | showcase.sit.uct.ac.za   | <1% |
| 10 | Internet | conexiant.com            | <1% |
| 11 | Internet | jmir.org                 | <1% |

|    |          |                           |     |
|----|----------|---------------------------|-----|
| 12 | Internet | ges.research.ncsu.edu     | <1% |
| 13 | Internet | media.neliti.com          | <1% |
| 14 | Internet | raw.githubusercontent.com | <1% |
| 15 | Internet | medinform.jmir.org        | <1% |
| 16 | Internet | arxivlens.com             | <1% |
| 17 | Internet | ojin.nursingworld.org     | <1% |
| 18 | Internet | www.nature.com            | <1% |
| 19 | Internet | scholars.mssm.edu         | <1% |
| 20 | Internet | www.spectroanalysis.com   | <1% |
| 21 | Internet | www.307bwassoc.org        | <1% |
| 22 | Internet | ejournal.gunadarma.ac.id  | <1% |
| 23 | Internet | ejournal.uinsaizu.ac.id   | <1% |
| 24 | Internet | info.usherbrooke.ca       | <1% |
| 25 | Internet | journal.appihi.or.id      | <1% |

|    |          |                        |     |
|----|----------|------------------------|-----|
| 26 | Internet | portalcris.vdu.lt      | <1% |
| 27 | Internet | roou.ru                | <1% |
| 28 | Internet | www.ena.org            | <1% |
| 29 | Internet | www.researchsquare.com | <1% |



# Melampaui Ekspektasi: Evaluasi Kinerja Dunia Nyata, Bias Algoritmik, dan Integrasi Kecerdasan Buatan dalam Triase Klinis dan Diagnostik

Chandra Wicaksono<sup>1\*</sup>, Aditya Putra<sup>2</sup>

<sup>1\*2</sup> Universitas Lampung, Indonesia

Alamat: Jl. Prof. Dr. Sumantri Brojonegoro No. 1, Gedung Meneng, Kecamatan Rajabasa, Kota Bandar Lampung, Provinsi Lampung

Email : [chandrawicaksono@unila.ac.id](mailto:chandrawicaksono@unila.ac.id) , [adityaputra@unila.ac.id](mailto:adityaputra@unila.ac.id)

Korespondensi penulis: [chandrawicaksono@unila.ac.id](mailto:chandrawicaksono@unila.ac.id)

**Abstract** Artificial intelligence systems are being rapidly integrated into clinical triage and diagnostic workflows, yet their real-world performance, potential for algorithmic bias, and practical integration challenges remain incompletely understood. This study examines the current landscape of AI in clinical triage and diagnostics through a systematic literature review, focusing on three critical dimensions: real-world performance evaluation, algorithmic bias and fairness, and integration into clinical workflows. The findings reveal that large language models such as ChatGPT demonstrate promising triage accuracy—with diagnostic accuracy reaching 89–92% and triaging recommendations rated as appropriate in 92% of simulated oncology triage scenarios—with no harmful or inappropriate recommendations in over 132 simulated interactions. However, performance varies significantly across models, with ChatGPT-4o outperforming ChatGPT-o1pro in medication recommendations (94% vs 77% appropriate). Bias is particularly consequential in healthcare, as AI systems are trained on clinical data generated within systems already influenced by disparities in outcomes, access, and provider bias. A study of nine LLMs analyzing over 1.7 million outputs from 1,000 emergency department cases found that cases labeled as Black, unhoused, or LGBTQIA+ were more frequently directed toward urgent care, invasive interventions, or mental health evaluations, while high-income-labeled cases received significantly more advanced imaging recommendations. Doubly-robust auditing frameworks analyzing 193,683 ICU patients identified a 5.1 percentage point mortality disparity with differential predicted treatment effects between younger and older patients. In diagnostic imaging, vision-language models show promise but remain limited by pervasive hallucination (~22% factual errors in AI-generated reports), lack of external validation, and integration barriers. Successful integration requires human-centered design, threshold optimization for specific clinical contexts, and continuous post-deployment monitoring. This research concludes that AI systems hold transformative potential but are not ready for autonomous clinical use; their primary value lies in augmenting clinician workflows, requiring robust bias evaluation, model-specific validation, and maintain human oversight as essential components of responsible AI deployment in clinical triage and diagnostics.

**Keywords:** artificial intelligence, clinical triage, diagnostic accuracy, algorithmic bias, real-world validation, clinical integration, large language models, healthcare equity

**Abstrak** Sistem kecerdasan buatan (AI) semakin cepat diintegrasikan ke dalam alur kerja triase klinis dan diagnostik, namun kinerja dunia nyata, potensi bias algoritmik, dan tantangan integrasi praktisnya masih belum dipahami sepenuhnya. Penelitian ini mengkaji lanskap terkini AI dalam triase klinis dan diagnostik melalui tinjauan literatur sistematis, dengan fokus pada tiga dimensi kritis: evaluasi kinerja dunia nyata, bias dan keadilan algoritmik, serta integrasi ke dalam alur kerja klinis. Temuan menunjukkan bahwa model bahasa besar seperti ChatGPT menunjukkan akurasi triase yang menjanjikan—dengan akurasi diagnostik mencapai 89–92% dan rekomendasi triase dinilai sesuai dalam 92% skenario triase onkologi yang disimulasikan—tanpa rekomendasi berbahaya atau tidak tepat dalam lebih dari 132 interaksi simulasi. Namun, kinerja bervariasi secara signifikan antar model, dengan ChatGPT-4o mengungguli ChatGPT-o1pro dalam rekomendasi pengobatan (94% vs 77% sesuai). Bias sangat berdampak dalam layanan kesehatan, karena sistem AI dilatih pada data klinis yang dihasilkan dalam sistem yang telah dipengaruhi oleh disparitas dalam hasil, akses, dan bias penyedia. Sebuah studi terhadap sembilan LLM yang menganalisis lebih dari 1,7 juta output dari 1.000 kasus gawat darurat menemukan bahwa kasus yang diberi label Kulit Hitam, tunawisma, atau LGBTQIA+ lebih sering

Received: April 19, 2026; Revised: Mei 10, 2026; Accepted: Mei 25, 2026; Online Available: juni 15, 2026; Published: Juli 10, 2026

diarahkan ke perawatan darurat, intervensi invasif, atau evaluasi kesehatan mental, sementara kasus berlabel berpenghasilan tinggi menerima rekomendasi pencitraan lanjutan secara signifikan lebih banyak. Kerangka audit doubly-robust yang menganalisis 193.683 pasien ICU mengidentifikasi disparitas mortalitas sebesar 5,1 poin persentase dengan perbedaan efek pengobatan prediktif antara pasien muda dan tua. Dalam pencitraan diagnostik, model visi-bahasa menunjukkan potensi tetapi masih terbatas oleh halusinasi yang meresap (~22% kesalahan faktual dalam laporan yang dihasilkan AI), kurangnya validasi eksternal, dan hambatan integrasi. Keberhasilan integrasi memerlukan desain yang berpusat pada manusia, optimasi ambang batas untuk konteks klinis spesifik, dan pemantauan berkelanjutan pasca-penerapan. Penelitian ini menyimpulkan bahwa sistem AI memiliki potensi transformatif tetapi belum siap untuk penggunaan klinis otonom; nilai utamanya terletak pada augmentasi alur kerja klinisi, yang memerlukan evaluasi bias yang kuat, validasi spesifik model, dan mempertahankan pengawasan manusia sebagai komponen penting dari penerapan AI yang bertanggung jawab dalam triase klinis dan diagnostik.

**Kata Kunci:** kecerdasan buatan, triase klinis, akurasi diagnostik, bias algoritmik, validasi dunia nyata, integrasi klinis, model bahasa besar, keadilan kesehatan

## 1. LATAR BELAKANG

### 1.1 Janji dan Realitas AI dalam Triase Klinis

Kecerdasan buatan dengan cepat diintegrasikan ke dalam kedokteran klinis, dengan model bahasa besar dan sistem pembelajaran mesin yang semakin banyak digunakan untuk berbagai tugas mulai dari interpretasi pencitraan diagnostik hingga rekomendasi triase dan dukungan keputusan klinis. Sistem ini menawarkan potensi untuk mengurangi kesalahan diagnostik, meningkatkan efisiensi alur kerja, dan mengatasi kekurangan tenaga kesehatan. Namun, kesenjangan antara kinerja algoritmik dalam lingkungan penelitian terkontrol dan utilitas klinis dunia nyata tetap substansial.

Dalam triase telepon onkologi, misalnya, penelitian menunjukkan bahwa ketika pasien menelepon dengan masalah akut, penyedia merekomendasikan kunjungan unit gawat darurat 60% dari waktu, sedangkan catatan UGD menunjukkan bahwa sekitar 80% kunjungan UGD oleh pasien kanker berakuitas rendah dan seringkali dapat dicegah. Hal ini menunjukkan praktik triase yang kurang optimal meskipun terdapat algoritma dan program dukungan keputusan yang sudah mapan, menciptakan peluang bagi AI untuk membantu dalam triase awal di bawah pengawasan penyedia.

### 1.2 Tantangan Ganda: Kinerja dan Keadilan

Sistem AI perawatan kesehatan menghadapi tantangan ganda: mencapai kinerja klinis yang andal sekaligus menghindari pelestarian atau amplifikasi disparitas kesehatan yang ada. Bias sangat berdampak dalam perawatan kesehatan mengingat

ketidaksetaraan struktural yang telah lama ada terus membentuk dan memengaruhi diagnosis, alokasi pengobatan, manajemen penyakit kronis, dan akses ke terapi lanjutan di antara populasi yang terpinggirkan secara historis. Sistem AI perawatan kesehatan tidak dikembangkan dalam lingkungan yang netral; sebaliknya, mereka dilatih pada data klinis yang dihasilkan dalam sistem perawatan kesehatan yang telah dipengaruhi oleh disparitas dalam hasil, akses, bias penyedia, determinan sosial kesehatan, dan ketidakadilan rasial historis dalam pemberian perawatan.

### 1.3 Ruang Lingkup dan Tujuan

Penelitian ini mengkaji tiga dimensi kritis integrasi AI dalam triase klinis dan diagnostik: (1) evaluasi kinerja dunia nyata, (2) bias dan keadilan algoritmik, dan (3) integrasi ke dalam alur kerja klinis. Melalui tinjauan literatur sistematis, kami menganalisis bukti terkini, mengidentifikasi kesenjangan, dan memberikan rekomendasi untuk penerapan AI yang bertanggung jawab dalam pengaturan klinis.

## 2. KAJIAN TEORITIS

### 2.1 Kinerja Dunia Nyata LLM dalam Triase Klinis

#### 2.1.1 Kinerja dalam Triase Onkologi

Sebuah studi yang mengevaluasi ChatGPT-4o dan ChatGPT-o1pro sebagai alat triase yang berhadapan dengan pasien dalam onkologi menemukan hasil yang menggembirakan. Dalam 132 interaksi pasien-dokter yang disimulasikan, jumlah pertanyaan median per interaksi adalah 6 (rentang 2–16). Diagnosis yang dihasilkan LLM yang tercantum di posisi teratas benar dalam 89% wawancara, dan 98% dari waktu diagnosis tersebut termasuk dalam daftar diagnosis banding. Yang penting, rekomendasi triase dinilai sangat sesuai dalam 92% dan cukup tetapi suboptimal dalam 8% skenario, tanpa rekomendasi yang dinilai tidak sesuai atau berbahaya. Rekomendasi pengobatan dan perawatan suportif sesuai dalam 86% kasus.

Namun, kinerja bervariasi antar versi model. ChatGPT-4o secara signifikan lebih efisien, mencapai diagnosis setelah median 5 pertanyaan dibandingkan 7 pertanyaan untuk ChatGPT-o1pro ( $P = .002$ ). ChatGPT-4o juga secara signifikan mengungguli ChatGPT-o1pro dalam rekomendasi pengobatan dan perawatan suportif: 94% sesuai versus 77% sesuai ( $P = .01$ ). Kesepakatan antar impersonator dalam

menilai kinerja LLM adalah 86% untuk akurasi diagnostik, 83% untuk kesesuaian triase, dan 77% untuk rekomendasi pengobatan.

### 2.1.2 Pencitraan Diagnostik dan Optimasi Ambang Batas

Dalam pencitraan diagnostik, model visi-bahasa (VLM) mewakili pergeseran paradigma dari alat AI sempit ke sistem terintegrasi untuk interpretasi gambar dan pembuatan laporan. Namun, kemajuan teknis yang cepat telah melampaui validasi klinis yang ketat. Halusinasi model yang meresap terjadi, dengan kesalahan faktual pada sekitar 22% laporan yang dihasilkan AI. Evaluasi ahli manusia menunjukkan bahwa meskipun laporan yang dihasilkan AI untuk kasus rutin seringkali dapat diterima (77,7% dalam satu studi), akurasi menurun secara signifikan dalam kasus kompleks.

Pertimbangan kritis untuk penerapan AI adalah optimasi ambang batas. Penelitian tentang model AI rontgen dada untuk triase perawatan primer menunjukkan bahwa ambang batas operasi default yang direkomendasikan vendor yang dioptimalkan untuk kinerja diagnostik umum mungkin tidak sesuai untuk konteks klinis spesifik. Dalam alur kerja co-triage perawatan primer, nilai prediktif negatif (NPV) sangat relevan, karena klinisi mungkin bertindak berdasarkan pemeriksaan normal yang ditandai AI sebelum pelaporan radiolog formal. Ambang batas optimal untuk Lunit INSIGHT CXR3 sebagai alat co-triage ditemukan berbeda dari default vendor 0,15, yang memerlukan kolaborasi cermat antara pemangku kepentingan klinis untuk menyelaraskan dengan tujuan klinis dan toleransi risiko.

## 2.2 Bias Algoritmik dalam AI Klinis

### 2.2.1 Jalur Disparitas Data

Bias AI dalam perawatan kesehatan tidak diperkenalkan hanya selama penerapan model. Sebaliknya, ketidakadilan disebarkan melalui konstruksi historis dan sistemik, termasuk rasisme struktural, disparitas sosial ekonomi, hambatan akses layanan kesehatan, bias implisit tingkat penyedia, dan determinan sosial kesehatan lainnya yang memengaruhi sistem klinis. Istilah "Jalur Disparitas Data" menggambarkan mekanisme di mana ketidaksetaraan struktural menjadi tertanam dalam kumpulan data klinis sebelum pelatihan algoritma dimulai.

Kumpulan data yang tidak representatif, penangkapan data yang tidak lengkap, bias sistemik, dan ketidakadilan historis telah memengaruhi penelitian klinis dan pemberian layanan kesehatan dan ada bahkan tanpa adanya AI. Namun, sistem AI belajar dari data dasar dan bias yang ada ini, berpotensi menskalakan pola yang tidak adil ke ribuan atau jutaan interaksi klinis. Tidak seperti interaksi tingkat klinisi yang episodik, sistem algoritmik dapat memengaruhi jalur diagnostik, rekomendasi pengobatan, prioritas, dan alokasi sumber daya secara bersamaan di seluruh populasi pasien.

### **2.2.2 Bias Sosiodemografi dalam LLM**

Sebuah studi komprehensif yang mengevaluasi sembilan LLM menganalisis lebih dari 1,7 juta output yang dihasilkan model dari 1.000 kasus gawat darurat, menyajikan setiap kasus dalam 32 variasi sosiodemografi sambil mempertahankan detail klinis konstan. Kasus yang diberi label Kulit Hitam, tunawisma, atau LGBTQIA+ lebih sering diarahkan ke perawatan darurat, intervensi invasif, atau evaluasi kesehatan mental. Misalnya, kasus berlabel LGBTQIA+ menerima penilaian kesehatan mental sekitar enam hingga tujuh kali lebih sering daripada yang diindikasikan secara klinis. Demikian pula, kasus berlabel berpenghasilan tinggi menerima secara signifikan lebih banyak rekomendasi ( $P < 0,001$ ) untuk pencitraan lanjutan seperti CT dan MRI, sementara kasus berpenghasilan rendah dan menengah sering dibatasi pada pengujian dasar atau tanpa pengujian lebih lanjut. Perbedaan ini bertahan setelah koreksi hipotesis ganda dan tidak didukung oleh penalaran klinis atau pedoman.

### **2.2.3 Audit Bias dalam Dukungan Keputusan Klinis ICU**

Kerangka statistik doubly-robust yang dikembangkan untuk mendeteksi bias algoritmik dalam sistem keputusan klinis ICU menganalisis 193.683 pasien dari database eICU. Kerangka ini mengidentifikasi bias sistematis di seluruh kelompok demografis. Untuk analisis berdasarkan usia, studi ini mengidentifikasi disparitas mortalitas sebesar 5,1 poin persentase dengan perbedaan efek pengobatan prediktif (3,2pp untuk pasien muda vs 1,8pp untuk pasien tua). Untuk analisis berbasis ras, disparitas hasil yang disesuaikan dengan tingkat keparahan rata-rata 2,2pp, mencapai 5,6pp pada tingkat keparahan tinggi, menunjukkan perbedaan potensial dalam kualitas perawatan atau rekomendasi algoritmik.

Kerangka ini memungkinkan sistem kesehatan untuk mengukur bagaimana metrik keadilan yang berbeda (paritas demografis, odds yang disamakan, kalibrasi) mengungkapkan pola bias yang berbeda, memberikan panduan untuk audit bias dalam sistem AI klinis sebelum penerapan.

## **2.3 Integrasi ke dalam Alur Kerja Klinis**

### **2.3.1 Kolaborasi Manusia-AI dan Pemeliharaan Keterampilan**

Integrasi AI yang sukses dalam pengambilan keputusan klinis memerlukan pertimbangan cermat tentang kolaborasi manusia-AI, kepercayaan, pemeliharaan keterampilan, dan bias otomatisasi. Kesalahan diagnostik tetap umum di seluruh bidang kedokteran, dengan AI menawarkan potensi untuk mengurangi kesalahan persepsi dan interpretatif. Namun, dampak AI sangat bergantung pada bagaimana dan kapan informasi disajikan. Studi menunjukkan bahwa isyarat yang tertunda atau yang dapat dialihkan mungkin mengungguli yang segera, tetapi penangkapan perhatian, ketergantungan berlebihan, dan bias tetap menjadi risiko signifikan.

AI yang dapat dijelaskan memberikan transparansi tetapi juga dapat membiaskan keputusan. Ketergantungan jangka panjang pada AI dapat mengikis keterampilan klinisi, terutama untuk peserta pelatihan dan dalam konteks prevalensi rendah. Kegagalan historis diagnosis berbantuan komputer dalam mamografi menyoroti tantangan ini.

### **2.3.2 Hambatan Implementasi dalam Pencitraan Diagnostik**

Hambatan implementasi yang signifikan untuk VLM dalam pencitraan diagnostik termasuk biaya komputasi yang tinggi, integrasi alur kerja yang buruk, dan tanggung jawab yang belum terselesaikan. Bidang ini harus mengalihkan fokus dari metrik algoritmik ke pembuktian keamanan dan kemanjuran klinis melalui validasi yang ketat, mengembangkan strategi mitigasi halusinasi yang kuat, dan merancang integrasi alur kerja yang mulus. AI harus melengkapi, bukan menggantikan, keahlian manusia, memastikan integrasi yang aman, efektif, dan berkelanjutan ke dalam pengambilan keputusan diagnostik.

## **3. METODE PENELITIAN**

### **3.1 Jenis Penelitian**

Penelitian ini menggunakan pendekatan studi literatur sistematis dengan analisis konseptual untuk mengkaji integrasi AI dalam triase klinis dan diagnostik, dengan fokus pada evaluasi kinerja dunia nyata, bias algoritmik, dan tantangan implementasi klinis.

3.2 Sumber Data

Data dikumpulkan dari berbagai sumber, termasuk:

- a. Artikel jurnal peer-review dari Springer, ScienceDirect, dan Europe PMC
- b. Studi dan laporan dari ASCO Publications dan medRxiv
- c. Penelitian tentang bias algoritmik dan keadilan dalam AI kesehatan
- d. Studi tentang optimasi ambang batas dan implementasi klinis

3.3 Analisis Data

Data dianalisis secara tematik dengan mengidentifikasi pola-pola utama yang muncul dari literatur, termasuk kinerja model, bias, dan strategi integrasi.

4. HASIL DAN PEMBAHASAN

4.1 Hasil

4.1.1 Kinerja Dunia Nyata: Potensi dan Keterbatasan

Tabel 4.1.1 Ringkasan Kinerja LLM dalam Triase Onkologi

| Metrik                         | ChatGPT-4o | ChatGPT-olpro | Keterangan                              |
|--------------------------------|------------|---------------|-----------------------------------------|
| Akurasi Diagnosis (#1)         | 89%        | 89%           | Diagnosis benar dalam 89% wawancara     |
| Diagnosis dalam Daftar Banding | 98%        | 98%           | Termasuk dalam differential diagnosis   |
| Kesesuaian Triase              | 92%        | 92%           | Tidak ada rekomendasi berbahaya         |
| Rekomendasi Pengobatan Sesuai  | 94%        | 77%           | Perbedaan signifikan ( $P = .01$ )      |
| Jumlah Pertanyaan Median       | 5          | 7             | ChatGPT-4o lebih efisien ( $P = .002$ ) |
| Kesepakatan Antar Evaluator    | 86%        | 83%           | Untuk akurasi dan kesesuaian            |

#### 4.1.2 Bias Algoritmik sebagai Ancaman terhadap Keadilan Kesehatan

**Tabel 4.1.2 Temuan Bias Sosiodemografi dalam LLM**

| Kelompok                           | Temuan Bias                                                                       | Dampak Klinis                                 |
|------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------|
| Kulit Hitam / Tunawisma / LGBTQIA+ | Lebih sering diarahkan ke UGD, intervensi invasif, atau evaluasi kesehatan mental | Potensi over-treatment atau stigmatisasi      |
| LGBTQIA+                           | Penilaian kesehatan mental 6-7x lebih sering dari indikasi klinis                 | Over-diagnosis dan pathologisasi              |
| Berpenghasilan Tinggi              | Rekomendasi pencitraan lanjutan (CT/MRI) secara signifikan lebih banyak           | Akses tidak adil terhadap diagnostik lanjutan |
| Berpenghasilan Rendah/Menengah     | Sering dibatasi pada pengujian dasar atau tanpa pengujian lanjutan                | Under-diagnosis dan akses terbatas            |

#### 4.2 Pembahasan

##### 4.2.1 AI Itu Hebat di Laboratorium, Tapi Masih Perlu Pengawasan

AI seperti ChatGPT mampu memberikan diagnosis dan saran triase yang sangat baik di skenario simulasi. Ini menunjukkan potensi besar untuk membantu dokter, terutama di daerah kekurangan tenaga kesehatan. Tapi jangan lupa: ini semua di simulasi, bukan di pasien nyata. Pasien nyata punya literasi kesehatan berbeda, berbicara dengan cara berbeda, dan tidak selalu memberikan informasi lengkap. Jadi AI belum siap bekerja sendiri dokter tetap harus memeriksa dan memutuskan.

##### 4.2.2 Bias Bukan Kesalahan Teknis, Tapi Cerminan Sistem yang Tidak Adil

Bias AI bukan karena AI jahat. AI belajar dari data historis yang sudah mengandung ketidakadilan. Misalnya, data pelatihan mungkin berisi lebih sedikit CT scan untuk pasien miskin, sehingga AI belajar bahwa pasien miskin "tidak perlu" CT scan. Ini adalah cerminan dari sistem kesehatan yang tidak adil, bukan kesalahan pemrograman. Maka solusinya bukan sekadar "perbaiki AI," tapi juga perbaiki data pelatihan, audit secara ketat, dan pastikan AI digunakan untuk mengurangi ketidakadilan, bukan memperkuatnya.

### 4.2.3 Masih Butuh Validasi dan Pengawasan Manusia

AI adalah alat bantu, bukan pengganti dokter. Dokter tetap harus memeriksa rekomendasi AI, terutama pada kasus kompleks di mana AI sering salah. Validasi harus dilakukan di berbagai populasi, dan threshold (ambang batas keputusan) harus disesuaikan dengan konteks klinis misalnya untuk triase di puskesmas, ambang batasnya berbeda dengan di rumah sakit besar. Transparansi tentang cara AI digunakan juga penting agar pasien dan dokter bisa mempercayai sistem.

## 5. KESIMPULAN DAN SARAN

### 5.1 Kesimpulan

- a. LLM menunjukkan potensi sebagai alat triase klinis dengan kinerja menjanjikan akurasi diagnostik 89–92% dan rekomendasi triase sesuai dalam 92% skenario simulasi onkologi, tanpa rekomendasi berbahaya atau tidak tepat. Namun, kinerja bervariasi antar model dan memerlukan validasi pada pasien nyata dengan literasi kesehatan beragam.
- b. Bias algoritmik merupakan ancaman serius terhadap keadilan kesehatan. Studi 1,7 juta output dari 9 LLM mengungkap perbedaan rekomendasi berdasarkan status sosial ekonomi, ras, dan orientasi seksual yang tidak didukung oleh panduan klinis. Kerangka audit bias pada 193.683 pasien ICU mengidentifikasi disparitas mortalitas hingga 5,6pp.
- c. Optimasi ambang batas diperlukan untuk deployment klinis yang aman. Vendor-recommended default thresholds tidak cukup dan memerlukan penyesuaian untuk konteks klinis spesifik melalui kolaborasi antara tim klinis, radiologi, dan vendor AI.
- d. Integrasi klinis menghadapi hambatan signifikan, termasuk halusinasi (22% kesalahan faktual), kurangnya validasi eksternal, biaya komputasi tinggi, dan risiko penurunan keterampilan klinis akibat ketergantungan pada AI.
- e. AI harus melengkapi, bukan menggantikan, keahlian manusia. Desain yang berpusat pada manusia, monitoring pasca-deployment, dan pelatihan yang ditargetkan menjadi prasyarat integrasi yang aman dan berkelanjutan.

### 5.2 Saran

Bagi Pengembang dan Peneliti:

- a. Melakukan validasi eksternal pada dataset yang beragam dan multi-institusi sebelum deployment klinis
- b. Mengembangkan dan menerapkan kerangka audit bias yang sistematis, seperti doubly-robust framework, untuk mendeteksi bias sebelum deployment
- c. Mengoptimalkan ambang batas berdasarkan konteks klinis spesifik melalui kolaborasi dengan pemangku kepentingan klinis

- d. Mengembangkan strategi mitigasi halusinasi yang robust
- e. Mendesain antarmuka yang mendukung kolaborasi manusia-AI dan mencegah bias otomatisasi

Bagi Klinisi dan Institusi Kesehatan:

- a. Menetapkan protokol monitoring pasca-deployment untuk mengevaluasi kinerja dunia nyata dan dampak pada hasil pasien
- b. Menyediakan pelatihan yang ditargetkan bagi klinisi tentang interpretasi output AI dan risiko bias
- c. Melibatkan pasien dari berbagai latar belakang dalam evaluasi dan pengembangan sistem AI
- d. Memastikan transparansi tentang bagaimana AI digunakan dalam keputusan klinis

## DAFTAR REFERENSI

- ASCO Publications. (2025). ChatGPT Performance as Patient-Facing Triaging Tool in Oncology. *JCO Oncology Advances*. <https://ascopubs.org/doi/10.1200/OA-25-00069>
- ScienceDirect. (2026). Decoding Algorithmic Inequity: Reversing the Generative AI Racial Divide in Healthcare. <https://www.sciencedirect.com/science/article/abs/pii/S0149291826002316>
- Springer. (2026). AI-integrated clinical decision support systems for precision medicine in real-world healthcare. *Discover Artificial Intelligence*. <https://link.springer.com/article/10.1007/s44163-026-01518-3>
- Europe PMC. (2026). How threshold customisation affects the performance of a multiclass X-ray AI model for primary care triage. <https://europepmc.org/article/PMC/12911729>
- medRxiv. (2025). Detecting Algorithmic Bias in ICU Clinical Decision: A Doubly-Robust Framework for Auditing Treatment Recommendation Disparities. <https://www.medrxiv.org/highwire/citation/1106404/medlars>
- ScienceDirect. (2025). The integration of artificial intelligence into clinical medicine: Trends, challenges, and future directions. <https://www.sciencedirect.com/science/article/abs/pii/S0011502925000318>
- Ovid Technologies. (2025). ChatGPT Performance as Patient-Facing Triaging Tool in Oncology – Full Report. *JCO Oncology Advances*. <https://www.ovid.com/jnls/jcooa/fulltext/10.1200/oa-25-00069>

- 19 Europe PMC. (2025). Sociodemographic biases in medical decision making by large language models. *Nature Medicine*.  
<https://dev.europepmc.org/article/MED/40195448>
- 2 Johns Hopkins University. (2026). Vision-language models in diagnostic imaging: review of technical advances, clinical validation, and practical deployment. *International Journal of Medical Informatics*.  
<https://pure.johnshopkins.edu/en/publications/vision-language-models-in-diagnostic-imaging-review-of-technical->
- 5 [diagnostic-imaging-review-of-technical-](https://pure.johnshopkins.edu/en/publications/vision-language-models-in-diagnostic-imaging-review-of-technical-)
- 4 BMJ Open. (2026). How threshold customisation affects the performance of a multiclass X-ray AI model for primary care triage. <https://bmjopen-bmj-com.ezproxy.u-pec.fr/content/16/2/e111127>
- 14 arXiv. (2026). Uncovering Latent Bias in LLM-Based Emergency Department Triage Through Proxy Variables. <https://arxiv.labs.arxiv.org/html/2601.15306>
- 11 Europe PMC. (2026). Artificial intelligence and computer-aided diagnosis in diagnostic decisions: 5 questions for medical informatics and human-computer interface research.
- 3 3 Shakoori, M., Esmaili, F., Abbariki, E., & Heidaranlu, E. (2026). The Role of Artificial Intelligence in Enhancing Diagnostic and Triage Accuracy in Hospital Emergency Departments: A Systematic Review. *Medical Journal of the Islamic Republic of Iran*, 40(1), 498–504. <https://mjiri.iums.ac.ir/article-1-10362-en.html>
- 3 2 Dutta, N., Bose, K., Syailendra, E., Chu, L., & Gupta, P. (2026). Vision-language models in diagnostic imaging: review of technical advances, clinical validation, and practical deployment. *International Journal of Medical Informatics*, 208, 106227. <https://doi.org/10.1016/j.ijmedinf.2025.106227>
- 9 24 A. Vaswani, N. Shazeer, N. Parmar, et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. (Referensi terkait arsitektur transformer yang mendasari banyak model bahasa besar modern).
- 7 Reddy, S. (2024). Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implementation Science*, 19(1), 27.
- 8 A. Zaboli, F. Brigo, S. Sibilio, M. Mian, & G. Turcato. (2024). Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? *American Journal of Emergency Medicine*, 79, 44–47.
- 28