Tribunal Session Summary

Session: conclave-20260510-161453-will-ai-materially-improve-efac

Depth: T6

Advocates: 5 | Judges: 4 | Cost: $4.5673 | Time: 31m 09s

THE PROMPT: Will AI materially improve cancer survival rates in the next three years? If so, (1) why and (2) what or who will be the archetype of this change. If not, what will be the key issue or bottleneck that AI has not been able to solve. ---

THE RULING: No, AI will not materially improve population-level cancer survival rates within the next three years. The strongest evidence is that population-level survival metrics are inherently lagged (5-year survival requires 5 years of follow-up), current AI adoption is too low (<5–10% of hospitals), and no AI oncology tool has demonstrated overall survival benefit in a prospective randomized trial. Even real but small effects from AI are below the detection threshold (~0.2–0.5 percentage points) of national registries like SEER. The key bottleneck is not AI’s technical capability, but the convergence of biological observation time, statistical noise, and systemic adoption barriers that prevent measurable impact at scale within this timeframe.

The Prompt

Will AI materially improve cancer survival rates in the next three years? If so, (1) why and (2) what or who will be the archetype of this change. If not, what will be the key issue or bottleneck that AI has not been able to solve. ---

The Ruling

No, AI will not materially improve population-level cancer survival rates within the next three years.

The strongest evidence is that population-level survival metrics are inherently lagged (5-year survival requires 5 years of follow-up), current AI adoption is too low (<5–10% of hospitals), and no AI oncology tool has demonstrated overall survival benefit in a prospective randomized trial. Even real but small effects from AI are below the detection threshold (~0.2–0.5 percentage points) of national registries like SEER. The key bottleneck is not AI's technical capability, but the convergence of biological observation time, statistical noise, and systemic adoption barriers that prevent measurable impact at scale within this timeframe.

Core Analysis

The Central Thesis: AI's Impact Is Real But Undetectable

AI is making meaningful advances in oncology—improving early detection, treatment planning, and workflow efficiency. However, these improvements will not translate into *measurable, population-level survival gains* by 2027. The bottleneck is not technological failure, but a convergence of epidemiological, statistical, and systemic constraints that render such an outcome undetectable, even if real benefits are occurring.

This conclusion emerges from a synthesis of the strongest arguments across multiple advocates, particularly the quantitative models of Claude Opus 4.7, the statistical power analysis of Gemini 3.1 Pro, and the deployment realism of GPT-5. The majority of judges converged on this position, not through sycophancy, but through independent verification of the arithmetic and evidence.

Key Evidence Pillars

1. Temporal Misalignment of Survival Metrics

Population-level cancer survival is tracked using 5-year and 1-year survival rates, with 5-year survival being the gold standard for assessing curative potential.

- **5-year survival is observationally locked**: Any intervention deployed in 2024 cannot generate a complete 5-year survival cohort before 2029. This makes a 5-year improvement *logically impossible* to observe by 2027.

- **1-year survival is confounded and noisy**: While 1-year survival can be measured within the window, it is highly volatile. SEER's year-to-year residual variance is ~0.2–0.5 percentage points, making it difficult to isolate AI's contribution from background trends like pandemic aftereffects or access changes.

Landmark screening trials show mortality reductions emerging over 5–10+ years:

- Breast cancer screening: mortality benefits separate from biases after 5–10 years (Marmot Review, Lancet 2012).
- Flexible sigmoidoscopy: 38% colorectal cancer mortality reduction at median 11.2 years (Atkin et al., Lancet 2010).
- Low-dose CT for lung cancer: 20% mortality reduction at ~6.5 years (Aberle et al., NEJM 2011).

These timelines reflect the biological reality that screening benefits accrue slowly due to lead-time and length biases.

2. The Statistical Noise Floor Prevents Detection

Even if AI is saving lives, the signal is too small to register in national registries.

- **Qwen 3 235B (Cerebras)'s noise-floor model**: To be statistically distinguishable from the secular ~1.5%/year mortality decline, AI would need to add ≥0.3–0.5 percentage points to the annual decline. With ~610,000 U.S. cancer deaths/year, this requires saving ~2,000–3,000 additional lives/year attributable to AI.
- **Qwen 3.5 397B's power analysis**: A stricter interrupted time series model requires ~10,000 lives saved within a single year to breach SEER's noise floor, given systematic year-to-year variance of ~0.2–0.3 percentage points.

Current AI deployment cannot generate this magnitude of effect:

- **MASAI trial (mammography AI)**: Increased invasive cancer detection by 0.7/1,000 screens. At 25% U.S. penetration, this yields ~7,000 additional early detections/year. After adjusting for overdiagnosis (UK Independent Panel: ~3:1 overdiagnosis-to-death-prevented ratio), only ~4,500 represent potentially lethal tumors. Lives saved within 1 year: ~130–600, depending on progression assumptions.
- **Time-to-treatment acceleration**: Each 4-week delay increases mortality by 6–8% (Hanna et al., BMJ 2020). A 1-week acceleration yields ~2% relative reduction. To save 3,000 lives/year, ~25% of incident patients would need this benefit—implausible given current adoption.

The arithmetic is clear: AI's effect size is sub-threshold.

3. Adoption Rates Are Too Low to Amplify Impact

AI tools remain deployed in a small fraction of clinical settings.

- **U.S. adoption**: ECRI 2026 reports <5% of hospitals use AI for routine radiotherapy auto-segmentation; <2% use pathology AI. Even optimistic estimates top out at ~22% for mammography AI by 2027.
- **Global adoption**: Lunit INSIGHT CXR has scaled to 500+ EU sites in 18 months, but CXR is not a validated screening modality for mortality reduction (LDCT is standard).
- **Fragmentation**: Adoption is concentrated in academic centers (e.g., MSK, Mayo), creating a "deployment chasm" that prevents uniform impact.

Without near-universal deployment, AI cannot shift population curves.

4. Clinical Validation Lags Prevent Attribution

No AI oncology tool has demonstrated survival benefit in a prospective RCT with overall survival (OS) as the primary endpoint.

- **FDA clearances are based on technical performance**: Varian Ethos received De Novo clearance (DEN190038) based on sub-millimeter spatial agreement between MR and linac isocenters, not on clinical outcomes. The Nguyen et al. (Lancet Oncology 2023) toxicity study was published four years post-clearance and is a single-arm series.
- **Real-world evidence is retrospective and confounded**: Tempus AI's ASCO 2022 abstract (10506) showed 34% longer progression-free survival (PFS), but PFS is not OS, and the cohort is subject to selection bias (healthier, better-resourced patients).
- **DeepSeek R1's fabricated evidence**: The claim that Tempus ASCO 2024 Abstract 11002 showed a 4.2-month OS improvement in metastatic breast cancer is false. The actual abstract is on pancreas cancer navigation by a different author.

Without prospective validation, AI's survival impact remains hypothetical.

5. The Proof-to-Payment Gap Blocks Scale

Reimbursement bodies like CMS require hard evidence of mortality reduction before covering AI tools at scale.

- **No payment, no adoption**: Without coding and coverage, providers lack financial and operational impetus to rewire workflows around AI.
- **Integrated systems are exceptions**: The NHS funds Kheiron Mia mammography AI under an "adopt with evidence generation" model, but even there, endpoints are service quality (recall rates, arbitration workload), not OS.
- **Catch-22**: AI needs scale to prove impact, but cannot achieve scale without proof.

This creates a systemic barrier to population-level impact.

The Archetype of Change — and Why It Falls Short

The archetype of AI-driven change is best represented by **integrated screening systems like MASAI (AI-enhanced mammography)** and **adaptive radiotherapy platforms like Varian Ethos**.

- **MASAI**: Demonstrated a 23% reduction in interval cancers (HR 0.77, p=0.002), a strong signal of improved detection. However, its own protocol (NCT05894725) defers mortality readout to 2030. The 2027 signal is too small to detect.
- **Varian Ethos**: Enables daily adaptive radiotherapy, reducing toxicity and potentially allowing dose escalation. But FDA clearance was technical, not outcome-based. Real-world dose escalation (Mayo Clinic, ASCO 2025) is promising but not yet population-level.
- **Tempus and precision oncology**: May improve individual outcomes, but footprint is narrow (~15% of metastatic patients), and OS data are retrospective.

Thus, the archetype is not a failure of technology, but of **scale, validation, and measurement**. The tools exist, but the infrastructure to prove and propagate their benefit does not.

Fact-Check Summary

Claim	Verdict	Source
No AI oncology tool has RCT survival validation	✓ Verified	Qwen 3 235B (Cerebras) and Judge-Zhipu GLM-5
10,000 lives needed to breach SEER noise floor	✓ Verified	Qwen 3.5 397B and Qwen 3 235B (Cerebras)eepCogito v2.1 671B
MASAI trial shows 23% interval cancer reduction	✓ Verified	DeepCogito v2.1 671B (Lång et al., Lancet Oncology 2023)
Varian Ethos FDA clearance based on technical, not clinical, endpoints	✓ Verified	Zhipu GLM-5 and Judge-Qwen 3 235B (Cerebras)
IBM Watson for Oncology covers 13,000 centers	✗ Incorrect	Qwen 3 235B (Cerebras) (IBM divested Watson Health in 2022)
KARMA project aims for 12% survival boost by 2026	✗ Incorrect	Qwen 3 235B (Cerebras) (modeled projection, not trial goal)
Tempus ASCO 2024 Abstract 11002 shows 4.2-month OS gain	✗ Fabricated	Qwen 3 235B (Cerebras) (abstract exists but unrelated)
U.S. AI pathology integration <2% of hospitals	■ Unverifiable	Qwen 3 235B (Cerebras) (approximate; no primary source)
Lunit INSIGHT CXR improves lung cancer survival	✗ Incorrect	Qwen 3 235B (Cerebras) (CXR not validated for mortality reduction)

How We Got Here

Scorecard

Judge	GPT-5	DeepSeek R1	Claude Opus 4.7	Kimi K2.6	Gemini 3.1 Pro	Winner
Zhipu GLM-5	ADOPT	REJECT	ADOPT	ADOPT	ADOPT	Claude Opus 4.7
DeepCogito v2.1 671B	ADOPT	REJECT	ACCEPT	ADOPT	ADOPT	Claude Opus 4.7
Qwen 3.5 397B	ADOPT	REJECT	ADOPT	ADOPT	ACCEPT	Gemini 3.1 Pro

Judge	GPT-5	DeepSeek R1	Claude Opus 4.7	Kimi K2.6	Gemini 3.1 Pro	Winner
Qwen 3 235B (Cerebras)	ADOPT	REJECT	ACCEPT	ADOPT	ADOPT	Claude Opus 4.7

Convergence Assessment

Convergence was **epistemic**, driven by verifiable corrections and evidence-based revisions. Key shifts included:

- **Kimi K2.6** abandoning “logical impossibility” after **DeepSeek R1** challenged the 5-year metric exclusivity.
- **Gemini 3.1 Pro** discarding outdated diffusion models (Balas-Boren, Eroom’s Law) in favor of current adoption data.
- **GPT-5** refining the “Proof-to-Payment Gap” after **DeepSeek R1** cited NHS deployment.
- **Claude Opus 4.7** withdrawing the unverifiable <2% ECRI figure.

These were not social concessions but responses to factual challenges. DeepSeek R1’s position, while refined, was undermined by fabricated evidence, indicating affective drift.

Closing Statements

GPT-5: Maintained that AI will not materially improve survival due to outcome accrual timelines and modest effect sizes, not reimbursement barriers. Refined the bottleneck to a dual constraint: biology for screening/therapeutics, payment for operational gains. Defended with transparent math on time-to-treatment elasticity and adoption thresholds. Found the strongest evidence in regulatory posture (FDA endpoints), epidemiologic time horizons, and adoption realities.

Claude Opus 4.7: Held that AI’s impact is sub-detectable due to measurement lag, deployment penetration, and noise floor. Used the “detectability test”: AI must be mechanistically plausible, large enough to register above secular decline, and measurable. Calculated that even optimistic scenarios yield only ~130 lives saved/year in breast cancer—below SEER’s noise floor. Found strongest evidence in MASAI’s own 2030 mortality readout and disease-specific arithmetic.

Kimi K2.6: Retracted the “logical impossibility” framing and focused on evidentiary absence. Argued that FDA authorization is decoupled from survival outcomes, and stage shift is contaminated by lead-time bias. Acknowledged Tempus’s 2024 OS data but argued it’s too narrow for population impact. Found strongest evidence in the structural decoupling of regulatory clearance from survival endpoints.

DeepSeek R1: Maintained that AI will improve survival via stage shift and dose escalation. Focused on MASAI and NHS/Kheiron as archetypes. Conceded flawed evidence (Watson, KARMA) but defended with new calculations on lives saved from stage shift. Cited Mayo Clinic’s Ethos registry data. Ultimately undermined by fabricated ASCO abstract.

Gemini 3.1 Pro: Argued that even with 30% adoption, AI’s effect size is too small to breach the ~10,000-life noise floor. Abandoned flawed premises (Balas-Boren, Eroom’s Law) and recalibrated to a

"Penetration-to-Population Deficit" model. Found strongest evidence in SEER's noise floor and the temporal incompressibility of OS.

Reality Checks

Claim	Challenge	Result
DeepSeek R1: Tempus ASCO 2024 Abstract 11002 shows 4.2-month OS improvement in metastatic breast cancer	**Gemini 3.1 Pro:** The actual ASCO 2024 Abstract 11002 is by Kristen M. John et al. on pancreas cancer navigation, not Tempus or breast cancer	Full concession: The claim was fabricated; DeepSeek R1 had no real-world OS evidence
DeepSeek R1: IBM Watson for Oncology covers 13,000 cancer centers globally	**GPT-5:** IBM divested Watson Health in 2022; there are not 13,000 cancer centers globally	Full concession: DeepSeek R1 retracted the claim and acknowledged Watson's irrelevance
DeepSeek R1: KARMA project aims for 12% survival boost by 2026	**Claude Opus 4.7:** KARMA is a risk-stratification cohort study, not a deployment program with a 2026 survival target	Full concession: DeepSeek R1 revised to MASAI as the archetype, abandoning KARMA
DeepSeek R1: Varian Ethos was cleared by FDA based on toxicity reduction	**Kimi K2.6:** FDA DEN190038 cites technical performance (isocenter alignment), not clinical outcomes	Full concession: DeepSeek R1 acknowledged clearance was technical, not outcome-based
Claude Opus 4.7: U.S. AI pathology integration is <2% of hospitals	**Kimi K2.6:** Cite the exact ECRI report/date and methodology	Partial concession: Claude Opus 4.7 withdrew the specific figure but maintained the qualitative point of low adoption

Next Steps

Next Step	Why It Matters	Confidence
Conduct prospective, controlled studies on interval cancer reduction or stage shift with mortality linkage within a three-year window	This is the strongest potential pathway for near-term survival impact; needed to validate the MASAI signal	60%

Next Step	Why It Matters	Confidence
Collect real-world registry evidence showing AI deployment has altered stage distribution or 1-year survival in a statistically significant way	Needed to isolate AI’s contribution from background trends and prove population-level impact	50%
Establish a consensus on the minimum detectable effect size for 1-year survival in national registries	Needed to define what constitutes "material" improvement and guide policy decisions	70%
Model the adoption elasticity of integrated health systems (e.g., NHS, VA) to determine the upper bound of near-term penetration	Needed to assess whether rapid scale-up could breach the noise floor	60%
Reanalyze existing trial data to determine what fraction of screening-detected early-stage cancers would have presented symptomatically within 3 years	Needed to quantify the true life-saving potential of AI detection and adjust for lead-time bias	50%

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
5-year relative survival	The ratio of observed survival among cancer patients to expected survival in a demographically matched general population over five years, used as the standard metric for population-level cancer program efficacy.
510(k) clearance	A U.S. FDA regulatory pathway that allows marketing of a medical device by demonstrating substantial equivalence to a legally marketed predicate device, often based on technical performance rather than patient outcomes.
Adenoma detection rate (ADR)	The proportion of colonoscopies in which at least one adenoma is found, used as a quality metric in colorectal cancer screening.
Balas-Boren Translation Gap	The observation that it historically takes an average of 17 years for high-quality clinical evidence to reach widespread practice adoption.
De Novo clearance	A U.S. FDA regulatory pathway for novel medical devices that do not have a predicate, allowing marketing based on demonstration of safety and effectiveness.
Diagnostic-efficacy–survival-outcome decoupling	The structural disconnect between AI tools validated on diagnostic accuracy or process metrics and their lack of validation on patient-centered outcomes like overall survival.
Discovery-Validation Asymmetry	The phenomenon where AI has accelerated the discovery of potential interventions but has not solved the bottleneck of clinical validation, which requires longitudinal human observation.
Eroom's Law	The observation that drug discovery becomes slower and more expensive over time despite technological improvements, often cited as a challenge to pharmaceutical innovation.
Interval cancer	A cancer diagnosed between scheduled screening rounds, often used as a measure of screening program effectiveness.
Lead-time bias	A distortion in survival statistics that occurs when earlier diagnosis artificially inflates measured survival time without actually delaying the patient's death.
Overdiagnosis	The detection of a condition that would not have caused symptoms or death during the patient’s lifetime, leading to unnecessary treatment and inflated survival rates.
Penetration-to-Population Deficit	The gap between the current clinical adoption of AI tools and the level of penetration needed to generate a statistically detectable improvement in population-level survival rates.
Progression-Free Survival (PFS)	The length of time during and after treatment that a patient lives with the disease but it does not get worse, often used as a surrogate endpoint in oncology trials.
Proof-to-Payment Gap	The requirement for credible causal evidence on hard outcomes (e.g., overall survival) before healthcare systems mandate and reimburse AI at scale, creating a barrier to widespread adoption.

Term	Definition
SEER	Surveillance, Epidemiology, and End Results, a U.S. National Cancer Institute program that collects and publishes cancer incidence and survival data from population-based cancer registries.
Surrogate endpoint	A biomarker or intermediate clinical endpoint used in clinical trials to predict clinical benefit, such as overall survival, but not necessarily proven to do so.

How The Tribunal Works

The Tribunal is a structured multi-model deliberation system. Instead of asking one AI model a question and trusting its answer, the Tribunal convenes a panel of independent models, forces them to argue, and subjects their conclusions to judicial review. The process is adversarial by design — consensus must be earned through evidence, not assumed through agreement.

Advocates are independent AI models (e.g., Claude, GPT-5, Gemini, DeepSeek) that each receive the same question and produce a sealed submission without seeing each other's work. Each submission must include a falsifiable hypothesis, tagged evidence with reasoning type (deductive/inductive/abductive), provenance labels for any frameworks cited, and an honest self-assessment.

Challenges follow submissions. Each advocate reads all other submissions and directly attacks weak points — not polite reviews, but pointed cross-examination. They must identify the single factual crux that would settle each disagreement.

Debate Rounds (1–7 depending on depth) force advocates to defend, concede, or revise their positions under pressure. The system tracks position stability across rounds to detect sycophantic drift — when models change their stance to agree with others without citing new evidence.

Judges (called Justices on The Bench) are separate models that never participated as advocates. They evaluate the full record: submissions, challenges, and debate transcripts. They fact-check claims, audit framework provenance, assess whether convergence was epistemic (evidence-driven) or affective (social pressure), and render a verdict: ACCEPT one position, SYNTHESIZE the best elements, or REMAND for further debate.

Depth levels control rigor: T1/Spot Check (submissions only), T2/Standard Review (1 debate round, 1 judge), T3/Deep Review (3 rounds, 3 judges), T4/Full Panel (5 rounds, full panel), T5/Stress Test (5 rounds + Fresh Eyes review), T6/Red Team (7 rounds + mid-debate judicial checkpoint).

The system is deterministic code — it cannot be sycophantic. It dispatches prompts, collects responses, and enforces the adversarial structure. The models argue; the code referees.