Tribunal Session Summary

Session: conclave-20260513-130333-will-ai-materially-improve-1b14

Depth: T6

Advocates: 6 | Judges: 6 | Cost: $4.9947 | Time: 41m 58s

THE PROMPT: Will AI materially improve pancreatic cancer survival rates over the next five years? ---

THE RULING: Partially. AI will materially improve survival for patients in early-adopting health systems by enabling earlier detection, but it will not produce a statistically or clinically material improvement in *national* 5-year pancreatic cancer survival rates by 2030 due to limited adoption, epidemiological measurement lag, and the small magnitude of attributable gains relative to secular trends.

The Prompt

Will AI materially improve pancreatic cancer survival rates over the next five years? ---

The Ruling

Partially. AI will materially improve survival for patients in early-adopting health systems by enabling earlier detection, but it will not produce a statistically or clinically material improvement in *national* 5-year pancreatic cancer survival rates by 2030 due to limited adoption, epidemiological measurement lag, and the small magnitude of attributable gains relative to secular trends.

The strongest evidence is the convergence among advocates on a narrow range of expected national survival gains (0.8–1.5 percentage points), derived from validated AI detection tools like Mayo Clinic's REDMOD and Placido's EHR risk model. This gain is real but falls below the ≥3 percentage point threshold required to be distinguishable from the pre-existing secular trend (~0.6 pp/year). The main caveat is definitional: while a 1 pp gain may be "material" in public health terms (hundreds of additional survivors), it does not represent a measurable acceleration of the national survival curve attributable to AI.

This ruling reflects the **SYNTHESIZE** verdict of the majority of judges, who concluded that AI will produce system-level gains but not population-level improvements. DeepCogito v2.1 671B explicitly ruled SYNTHESIZE, while Qwen 3 235B (Cerebras) and Zhipu GLM-5.1 supported outcomes compatible with synthesis. Zhipu GLM-5's ACCEPT [Claude Opus 4.7] and Qwen 3 235B (Cerebras)'s REMAND are minority positions.

Core Analysis

The Dual-Level Impact of AI on Pancreatic Cancer Survival (2025–2030)

The tribunal's deliberation reveals a fundamental divergence between AI's *mechanistic potential* and its *population-level impact*. While AI tools are demonstrably capable of detecting pancreatic ductal adenocarcinoma (PDAC) earlier, the structural realities of healthcare deployment, epidemiological measurement, and the definition of "material improvement" prevent a measurable shift in national 5-year survival rates by 2030.

1. Mechanistic Pathways: How AI Can Improve Survival

AI is contributing to earlier PDAC detection through two primary, evidence-backed mechanisms:

- **Opportunistic CT-AI detection**: The Mayo Clinic's REDMOD AI model—verified in a *Gut* 2026 multi-institutional study—detects 73% of prediagnostic PDAC cases approximately 16 months before clinical diagnosis by analyzing routine abdominal CT scans. This represents a validated, scalable

pathway to stage-shift. The model operates as a software layer within existing PACS systems, flagging subtle textural and ductal changes that human radiologists often miss. Its median lead time of ~16 months is sufficient to move a subset of patients from metastatic (stage IV) to resectable (stage I/II) disease.

- **EHR-based risk stratification**: Models like the one developed by Placido et al. (*Nature Medicine*, 2023) achieve an AUROC of ~0.88 in predicting 3-year PDAC risk using structured and unstructured EHR data. When deployed, these tools can flag high-risk patients—such as those with new-onset diabetes after age 50—for targeted imaging (e.g., MRI or EUS), mimicking the successful CAPS5 surveillance protocol, which achieved a 73% 5-year survival rate in *surveillance-detected* cases (*JAMA*, 2022).

These tools are not speculative. They are in active use or advanced piloting at institutions like Mayo Clinic, Kaiser Permanente, and MD Anderson. The stage-shift arithmetic is sound: shifting even 4–6% of cases from stage IV (3% 5-year survival) to stage I/II (30–55% survival) yields a net gain of ~1,600–2,000 additional 5-year survivors nationally over five years.

2. Barriers to Population-Level Impact

Despite this mechanistic promise, **national 5-year survival rates will not show a material improvement by 2030** due to:

- **Limited deployment**: Current adoption of AI-enabled PDAC screening is confined to a small number of integrated delivery systems. DeepCogito v2.1 671B estimates 15% population coverage by 2030; others suggest lower. Without widespread implementation, the denominator effect dilutes AI's impact. The MedPAC 2025 report shows that 65% of community hospitals operate at <3% margins, making them unable to absorb the cost of AI software without dedicated reimbursement.
- **Reimbursement gaps**: No CPT code exists for AI-assisted cancer screening. Hospitals operate on thin margins, and without CMS reimbursement, scaling beyond research settings is economically unviable. This structural barrier is emphasized by Grok 4 and Gemini 3.1 Pro and supported by current FDA-CMS parallel review pilot data.
- **Measurement lag**: While period analysis allows for near-real-time survival estimates, the 5-year survival metric still reflects cohorts diagnosed years prior. AI tools only began meaningful deployment in 2025–2026; their full impact will not be measurable until 2031–2032. Gemini 3.1 Pro's "Chronological Catch-22" was overstated, but the core insight—**temporal misalignment between intervention and metric**—is valid.

As a result, projected national gains are modest: **0.8–1.5 percentage points** in 5-year survival. This is within the range of the secular trend (~0.6 pp/year, or ~3 pp over five years), driven by improvements in surgery, chemotherapy (e.g., FOLFIRINOX), and supportive care.

3. The Definition of "Material Improvement"

The central conflict is not empirical but **semantic and normative**. Advocates and judges disagree on what constitutes a "material" improvement:

- **Grok 4, Gemini 3.1 Pro, Claude Opus 4.7, and Kimi K2.6, and E** define "material" as ≥3–5 pp absolute gain—large enough to be statistically distinguishable from secular trends and clinically meaningful at the population level.
- **GPT-5** defines "material" as ≥1 pp attributable gain—even if within the secular trend—because it represents ~650 additional 5-year survivors annually.

No external, validated standard for “material” was provided. In oncology policy, a 5 pp improvement is often considered clinically significant, but public health frameworks may treat 1 pp as meaningful given PDAC’s lethality (baseline 5-year survival: ~12%).

The tribunal’s synthesis resolves this by **distinguishing system-level from population-level impact**:

- **In early-adopting systems**, AI will produce **material improvements** (2–3 pp gain, stage-shift of 4–6%, survival approaching 30–40% in detected cohorts).
- **At the national level**, gains will be **real but not material** under a ≥3 pp threshold, and **statistically inseparable** from background trends.

4. Attribution and Verification

The only path to causal attribution is through **difference-in-differences (DiD)** or **stepped-wedge designs** in early-adopting systems. Such studies are underway at Mayo and Kaiser and may yield results by 2028–2029. These will allow isolation of AI’s effect from concurrent improvements in treatment.

Until then, national statistics cannot reliably attribute survival gains to AI. The convergence among Grok 4, Gemini 3.1 Pro, Claude Opus 4.7, and Kimi K2.6, and F on a 0.8–1.5 pp gain—despite starting from different premises—lends credibility to this estimate.

5. Therapeutic AI Is Not a Factor (Yet)

All advocates agree: **no AI-discovered or AI-optimized therapeutic** will complete Phase III trials and reach patients by 2030. Claims by DeepSeek R1 regarding “PANC-1” (a cell line, not a drug) and Insilico’s pipeline were fabricated. AlphaFold2 aids structural biology but does not autonomously identify druggable targets. Thus, AI’s impact is confined to **diagnosis and triage**, not treatment.

How We Got Here

Scorecard

Judge	Grok 4	Gemini 3.1 Pro	Claude Opus 4.7	Kimi K2.6	DeepSeek R1	GPT-5	Winner
Qwen 3 235B (Cerebras)	ADOPT	ADOPT	ADOPT	ADOPT	REJECT	ADOPT	GPT-5
Qwen 3.5 397B	REJECT	ADOPT	ACCEPT	ADOPT	REJECT	ADOPT	Claude Opus 4.7
Zhipu GLM-5	REJECT	ADOPT	ACCEPT	ADOPT	REJECT	ADOPT	Claude Opus 4.7

Judge	Grok 4	Gemini 3.1 Pro	Claude Opus 4.7	Kimi K2.6	DeepSeek R1	GPT-5	Winner
Zhipu GLM-5.1	ADOPT	ADOPT	ADOPT	ADOPT	REJECT	ADOPT	Claude Opus 4.7
DeepCogito v2.1 671B	ADOPT	ADOPT	ADOPT	ADOPT	REJECT	ADOPT	Claude Opus 4.7
Kimi K2.6	ADOPT	ADOPT	ADOPT	ADOPT	REJECT	ADOPT	Claude Opus 4.7

Convergence Assessment

Convergence was **epistemic**, driven by verifiable corrections and evidence-based concessions. The collapse of DeepSeek R1's fabricated claims forced all advocates to rely on peer-reviewed studies. The concession of the "Chronological Catch-22" and the refinement of the stage-shift model were triggered by specific, factual challenges. Position stability data shows that advocates changed their views only when presented with new evidence, not due to social pressure.

Closing Statements

GPT-5 My final position is YES: AI will materially improve pancreatic cancer survival. I expect a 2–3 pp gain in early-adopting systems and 0.8–1.5 pp nationally by 2030, driven by opportunistic CT-AI (REDMOD) and EHR triage. I learned from the debate that EUS capacity is a real constraint, so I revised my model to include MRI-first gating, reducing incremental EUS to the low tens of thousands. My strongest evidence is the REDMOD study, which shows 73% detection ~16 months early, and the CAPS5 data, which proves that early detection yields 73% 5-year survival. I maintain that a 1 pp gain is material, even if within secular trends.

Grok 4 (DeepSeek R1) My final position is NO: AI will not materially improve survival (>3 pp) due to deployment and regulatory timelines. I conceded three fabricated citations (Park et al., Watson, PANCAIM) but rebuilt on verified sources like SEER 2024 data. I learned that AI's impact is real but small, so I refined my threshold to >3 pp. My strongest evidence is the FDA's 7.2-year average from Phase I to approval, the SEER trend of 0.5–1 pp per 5 years, and the JAMA Oncology meta-analysis showing no significant survival improvements from AI in oncology.

Gemini 3.1 Pro My final position is NO: AI will not materially improve population-level survival due to economic and systemic barriers. I abandoned the "Chronological Catch-22" after conceding period analysis, but pivoted to the economics of scale: without CPT codes, AI is restricted to <15% of the population. I learned from GPT-5 that MRI-first gating makes the workflow feasible, but this only confirms that the stage shift is too small to matter nationally. My strongest evidence is the math: a 3% stage shift yields only 1.35 pp survival gain, less than half the secular trend of 0.6 pp/year.

Claude Opus 4.7 My final position is NO: AI will not produce ≥5 pp improvement in national survival by 2030. I conceded that stage I survival is higher (~44–55%) and that deployment is not "zero," but maintain that

0.8–1.5 pp is below materiality. I learned from GPT-5 that DiD designs can show system-level gains, but these do not translate to the population. My strongest evidence is the structural lag of 5-year survival metrics, the historical base rate of AI in oncology (1 pp gain in breast cancer), and the biological ceiling on stage-shift alone.

DeepSeek R1 My final position is NO: AI will not produce a materially attributable improvement in national survival by 2030. I abandoned my YES position after conceding that my key citations (PancreasPlus-AI, PANC-1, Tempus) were fabricated or incorrect. I learned that evidence integrity is paramount. My strongest surviving evidence is the coverage dilution math: 15% adoption limits the national gain to 0.6 pp, below statistical detectability.

Reality Checks

Claim	Challenge	Result
DeepSeek R1: "Insilico's AI-designed pancreatic cancer inhibitor (PANC-1) entered Phase I trials in 2023"	**Claude Opus 4.7:** "PANC-1 is a well-known pancreatic cancer cell line, not an Insilico compound. Insilico's lead program is INS018_055 for idiopathic pulmonary fibrosis, not pancreatic cancer."	Full concession: DeepSeek R1 admitted PANC-1 is a cell line and withdrew the claim.
Grok 4: "IBM Watson Health discontinued in 2022 after trials showed no survival benefits"	**Claude Opus 4.7:** "Watson Health was sold to Francisco Partners in 2022, not 'discontinued after trials showed no survival benefits.' You are conflating a business divestiture with a clinical efficacy failure."	Full concession: Grok 4 admitted the mischaracterization and revised the claim.
Gemini 3.1 Pro: "It is mathematically impossible for a '5-year survival rate' metric to capture a material population-wide shift by 2031"	**Claude Opus 4.7:** "Period analysis and conditional survival methods routinely produce 5-year survival estimates before 5 years of follow-up has accrued... Your 'patient must literally survive until 2032' framing is wrong as stated."	Full concession: Gemini 3.1 Pro abandoned the "Chronological Catch-22" and revised the argument to focus on deployment economics.
GPT-5: "EHR-based DL triage achieves PPVs enabling viable surveillance"	**Claude Opus 4.7:** "PPV of ~3–5% in the top centiles means 95–97% false positives going to EUS/MRI workup. The harm/cost ratio of that pipeline at population scale is not benign. You hand-wave 'workable' — defend it with numbers."	Revision: GPT-5 revised the model to include MRI-first gating, reducing EUS burden and making the pathway more feasible.

Next Steps

Next Step	Why It Matters	Confidence
Define "material improvement" in cancer survival rates (e.g., ≥1 pp vs. ≥5 pp)	Without a shared definition, policy decisions lack a benchmark.	70%
Monitor DiD or stepped-wedge studies in early-adopting systems (e.g., Mayo, Kaiser) for causal evidence of AI impact	This is the only path to attribution before 2030.	70%
Track CMS decisions on CPT codes for AI-assisted cancer screening	Reimbursement is the key enabler for scaling beyond pilot programs.	30–50%
Validate generalizability of CAPS5 and REDMOD results in average-risk populations	Real-world effectiveness may be lower than in high-risk cohorts.	60%

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
5-year relative survival rate	The percentage of patients who survive at least 5 years after diagnosis, adjusted for age and other causes of death.
AUROC	Area Under the Receiver Operating Characteristic curve, a measure of a model's ability to distinguish between classes.
CAPS5	A surveillance program for high-risk individuals that achieved 73% 5-year survival in detected pancreatic cancers.
CPT code	Current Procedural Terminology code, used for medical billing and reimbursement in the US.
DiD	Difference-in-differences, a statistical method to estimate causal effects by comparing changes over time between treatment and control groups.
EHR	Electronic Health Record, a digital version of a patient's paper chart.
FDA	Food and Drug Administration, the US agency responsible for protecting public health by ensuring the safety of medical products.
MCED	Multi-cancer early detection, blood tests that use machine learning to detect cancer signals.
PACS	Picture Archiving and Communication System, used for storing and retrieving medical images.
PDAC	Pancreatic ductal adenocarcinoma, the most common type of pancreatic cancer.
PPV	Positive Predictive Value, the proportion of positive test results that are true positives.
SEER	Surveillance, Epidemiology, and End Results program, a source of cancer statistics in the US.
Stepped-wedge design	A type of clinical trial where all participants eventually receive the intervention, but the timing is randomized.

How The Conclave 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.