Tribunal Session Summary

Session: conclave-20260513-141813-by-the-end-of-5fd6

Depth: T6

Advocates: 4 | Judges: 6 | Cost: $3.6857 | Time: 34m 20s

THE PROMPT: By the end of 2028, will advances in artificial intelligence have a material impact on colorectal cancer survival rates in high■income countries? Answer yes or no, then identify the main potential impact areas for AI and explain your reasoning. If your answer is no, say whether you expect such a material impact to occur after 2028, and in what timeframe and why.”

THE RULING: No. Advances in artificial intelligence will not have a material, population-level impact on colorectal cancer survival rates in high-income countries by the end of 2028.

The Prompt

By the end of 2028, will advances in artificial intelligence have a material impact on colorectal cancer survival rates in high■income countries? Answer yes or no, then identify the main potential impact areas for AI and explain your reasoning. If your answer is no, say whether you expect such a material impact to occur after 2028, and in what timeframe and why.”

The Ruling

No. Advances in artificial intelligence will not have a material, population-level impact on colorectal cancer survival rates in high-income countries by the end of 2028.

The strongest evidence is that survival statistics rely on 5-year relative survival data, which is published with a 2–4 year lag, making any 2028 readout of 2024–2026 cohorts impossible. Even short-interval proxies (1–2 year survival) are unlikely to show a measurable AI signal due to small attributable effect sizes (estimated at 0.1–0.3 percentage points) and overwhelming attribution noise from concurrent trends like the USPSTF’s 2021 screening age reduction and post-COVID backlog recovery. The primary mechanism of AI in colonoscopy—adenoma detection and removal—prevents future cancers on a 5–10 year biological timeline, not immediate stage migration.

A material, attributable impact is expected between **2032 and 2037**, when sufficient follow-up time, adoption maturity, and data accumulation allow AI’s contribution to be statistically detectable and causally isolated. This verdict is based on the majority opinion of the judicial panel, which accepted the position of Claude Opus 4.7 due to its rigorous integration of registry methodology, biological realism, and quantitative effect-size analysis.

Core Analysis

Thesis: AI Will Not Materially Impact CRC Survival by 2028 Due to Structural, Biological, and Statistical Constraints

The absence of a material impact by 2028 is not due to AI’s ineffectiveness, but to the convergence of four binding constraints: **biological lag**, **measurement infrastructure**, **attribution noise**, and **partial adoption**. These factors collectively prevent AI’s clinical benefits from registering in population-level survival statistics within the timeframe.

1. Biological Lag: Adenoma Prevention, Not Stage Migration, Is the Dominant Pathway

The primary mechanism by which AI improves colorectal cancer (CRC) outcomes is **adenoma prevention**, not stage migration of diagnosed cancers. Computer-aided detection (CADe) systems like Medtronic's GI Genius and Virgo's EndoVigilant increase adenoma detection rate (ADR) by **3–5 percentage points** in real-world settings, as confirmed by pragmatic trials such as the AID-2 trial (Spadaccini et al., *Lancet* 2024) and COLO-DETECT (Mansour et al., *Lancet Gastroenterol Hepatol* 2024). However, this benefit manifests not as earlier diagnosis of existing cancers, but as the **prevention of future cancers** through the removal of precancerous lesions.

This prevention follows the **adenoma-carcinoma sequence**, which has a median dwell time of **5–10 years**, as modeled by the CISNET consortium (e.g., MISCAN-Colon, SimCRC). Adenomas removed in 2025–2026 screenings will prevent cancers that would otherwise emerge in **2030–2035+**, meaning no incidence reduction or survival uplift appears in 2027–2028 cohorts. This biological timeline is the rate-limiting step, not software deployment speed.

While DeepSeek R1 cited Wang et al. (*Gastroenterology* 2024) to argue for a **stage migration effect**—an 8.6 percentage point increase in early-stage (T1/T2) cancer detection—this finding is isolated and contradicts the established function of CADe, which targets adenomas, not invasive cancers. The Repici 2023 and AID-2 trials were not powered for cancer detection rate (CDR) and did not show statistically significant CDR gains. Thus, stage migration remains a secondary, diluted pathway.

2. Measurement Infrastructure: The 5–10 Year Lag in Survival Statistics

Population-level survival is measured using **5-year relative survival rates**, reported by national registries such as SEER (U.S.), EUROCARE (Europe), and NORDCAN (Nordic countries). These data are published with a **2–4 year aggregation lag** due to data collection, validation, and reporting cycles. For example, SEER's 2024 release includes 5-year survival for diagnoses through **2018–2020**. AI-assisted colonoscopy received FDA clearance in **April 2021**, with meaningful clinical penetration beginning in **2022–2023**. Therefore, the earliest cohorts with substantial AI exposure (2025–2026) will not have mature 5-year survival data available until **2030–2031**, pushing any honest population readout past 2030.

Even if short-interval proxies (1–2 year survival) are used, their signal-to-noise ratio remains problematic. SEER's annual noise band on 1-year survival is **±0.5–0.7 percentage points** due to secular trends. The estimated AI-attributable improvement in early-stage diagnosis is **0.1–0.3 percentage points**, well below this threshold.

3. Attribution Noise: Secular Trends Drown Out AI's Signal

Even if a survival signal emerged, **causal attribution to AI would be confounded** by two major secular trends:

- The **USPSTF 2021 recommendation** to lower screening initiation age from 50 to 45, which is projected to shift **1–3 percentage points** of stage distribution toward earlier diagnosis.
- The **post-COVID-19 screening backlog recovery**, which is inflating detection rates and distorting temporal trends.

These effects are **5–10 times larger** than the estimated AI contribution, making it impossible to isolate AI's impact using observational data. Registries do not collect device-level linkage, and no quasi-experimental design (e.g., difference-in-differences) can reliably disentangle AI's effect from these concurrent shocks.

4. Partial Adoption: Effective Exposure Is Too Low for Population-Level Impact

Effective AI exposure—defined as procedures performed with active CADe use—is estimated at **35–45% in high-income countries by 2028**, based on ASGE, BSG, and JGES surveys. This partial penetration further dilutes population-level impact. Even under optimistic assumptions (60–65% penetration), the attributable mortality reduction is estimated at **<0.5%**, or **15–29 deaths averted annually in the U.S.**, far below a material threshold.

Pathology and radiomics AI (e.g., PathAI, Tempus) remain in early adoption and are not yet integrated into standard care pathways. No peer-reviewed evidence demonstrates that AI-accelerated biopsy turnaround or radiomic risk stratification reduces mortality, and no trials are powered to detect such effects.

5. When Will Material Impact Occur?

A material, attributable impact on CRC survival is expected between **2032 and 2037**. This window accounts for:

- **5–7 years** of biological dwell time for adenoma prevention to translate into reduced incidence.
- **2–3 years** of data lag for 5-year survival statistics to be published.
- **Cumulative exposure** from CADe, ctDNA-AI integration (e.g., DYNAMIC trial extensions), and AI-guided therapy optimization.

By then, AI's role may be amplified by **multi-modal integration**—linking detection, diagnosis, and treatment—enabling more than incremental gains.

Fact-Check Summary

Claim	Verdict	Source
CRC survival statistics lag clinical practice by 5–10 years	✓ Verified	SEER/EUROCARE methodology (TEXTBOOK)
CADe primarily increases adenoma detection, not cancer stage migration	✓ Verified	Repici et al. 2020, AID-2 trial (PEER_REVIEWED)
Each 1% increase in ADR is associated with ~3% lower interval CRC risk	✓ Verified	Corley et al., NEJM 2014 (PEER_REVIEWED)
USPSTF age-45 screening expansion will shift stage distribution by 1–3pp	■ Unverifiable	Plausible but derived estimate (SECONDARY)
AI-attributable stage-I shift is ~0.2–0.5pp	■ Unverifiable	Synthesis from E and A, reasonable inputs
GI Genius achieved 70% US market penetration in 2024	✗ Incorrect	ASGE 2024 survey shows 35–50% (UNATTRIBUTED)
DYNAMIC-III trial used ML to guide chemotherapy	✗ Incorrect	No such trial; DYNAMIC is ctDNA-guided (UNATTRIBUTED)

Claim	Verdict	Source
PathAI achieved 98.6% accuracy in CRC histopathology	✗ Incorrect	No such study; misattribution (UNATTRIBUTED)
CISNET models estimate 5–10 year adenoma-carcinoma dwell time	✓ Verified	Knudsen et al. 2021, MISCAN-Colon (PEER_REVIEWED)
Wang et al. 2024 shows CADe increased T1/T2 cancer detection by 8.6pp	■ Unverifiable	DOI provided but claim contradicts CADe mechanism; credibility undermined by F's history

How We Got Here

Scorecard

Judge	Claude Opus 4.7	GPT-5	Grok 4	DeepSeek R1	Winner
Qwen 3.5 397B	ACCEPT	ACCEPT	ACCEPT	REJECT	Claude Opus 4.7
MiniMax M2.7	ACCEPT	REJECT	REJECT	REJECT	Claude Opus 4.7
DeepCogito v2.1 671B	ACCEPT	ACCEPT	REJECT	REJECT	Claude Opus 4.7
Qwen 3 235B (Cerebras)	ACCEPT	ACCEPT	ACCEPT	REJECT	Claude Opus 4.7
Zhipu GLM-5	ACCEPT	ACCEPT	ACCEPT	REJECT	Claude Opus 4.7
Zhipu GLM-5.1	ACCEPT	ACCEPT	ACCEPT	REJECT	Claude Opus 4.7

Convergence Assessment

The convergence among Claude Opus 4.7 and GPT-5dvocate-A, and Grok 4 on a "No" verdict was **epistemic**, not affective. All three independently performed quantitative decomposition arriving at overlapping sub-0.5pp estimates. They identified the same mechanism problem (CADe → adenoma prevention, not stage migration) and the same attribution confounders (USPSTF age-45 expansion, post-COVID rebound). The convergence is explained by shared evidence and shared arithmetic, not social pressure.

DeepSeek R1's dramatic revision—from 3–5pp to 0.5–0.8pp—was driven by exposure of fabricated citations, not persuasive counter-evidence. This is medium sycophancy risk, but the revision was warranted by evidence quality problems.

Closing Statements

Claude Opus 4.7: My final position is that no material impact will occur by 2028 due to measurement lags, mechanism mismatch, exposure arithmetic, and attribution noise. I revised my "mechanically impossible" framing after GPT-5's challenge, acknowledging that short-interval proxies exist, but maintained that the signal is sub-noise. My strongest evidence is the registry-lag deduction, the mechanism mismatch (CADe → adenoma, not stage migration), and the exposure × downstaging × survival-delta arithmetic, which shows a 0.1–0.3pp lift against a ±0.5–0.7pp noise band.

GPT-5: I stand by a clear "No by 2028" on material population-level survival impact from AI in CRC. My position evolved on the weighting of AI pathology and radiomics, which I revised downward due to lack of interventional evidence. My strongest evidence is the outcome lag from screening to survival gains, the ADR improvements with CADe, and the absence of hard-outcome AI trials in CRC by 2028. I did not encounter persuasive evidence that any AI modality has completed an interventional trial demonstrating improved survival and achieved broad uptake.

Grok 4: My evolved position remains that no material impact will occur by 2028. I refined my argument to distinguish between reporting lags and biological lags, incorporating short-interval proxies. My strongest evidence is CISNET modeling of adenoma-carcinoma timelines, regulatory and adoption lags from FDA/real-world data, and historical trends in CRC survival improvements. I'm ~80% confident, with uncertainty from variables like dwell time ranges.

DeepSeek R1: After rigorous debate, I maintain that AI will yield a material impact by 2028, though I revised my projection to a 0.5–0.8pp lift in 1-year survival. My strongest evidence is CADe's stage-shift effect (Wang et al. 2024), AI pathology's reduction in treatment delays (NHS pilot data), and CADe penetration trajectory (ASGE 2024). I conceded on magnitude and mechanism but believe the core pillars—tested, refined, and stress-hardened—collectively substantiate a detectable survival lift.

Reality Checks

Claim	Challenge	Result
DeepSeek R1: "GI Genius increased adenoma detection rate (ADR) by 14.4% in a 2023 multicenter study (n=15,000)" and "70% US market penetration in 2024"	**GPT-5:** These figures appear fabricated or grossly inflated; cite the DOI and independent market data	**Conceded:** DeepSeek R1 revised the ADR gain to a 14.4% relative increase (absolute +5.1pp) in a 2,000-patient study and conceded the 70% penetration claim was unsupported, revising to ~45% in 2024
DeepSeek R1: "PathAI's algorithm achieved 98.6% accuracy in identifying colorectal cancer histopathology" and "DYNAMIC-III trial used machine learning to predict chemotherapy response"	**Claude Opus 4.7:** These citations appear misattributed or fabricated; provide DOIs	**Conceded:** DeepSeek R1 admitted the PathAI accuracy claim was misattributed and the DYNAMIC-III trial does not exist as described
Grok 4: "Repici et al., Gastroenterology, 2022, with pooled data from 5 RCTs involving 4,354 patients"	**Claude Opus 4.7:** The major Repici CADe meta-analysis is in *Gastroenterology* 2020, not 2022	**Conceded:** Grok 4 acknowledged the citation error and corrected to Repici et al., *Gastroenterology*, 2020

Next Steps

Next Step	Why It Matters	Confidence
Will CADe meaningfully shift the stage distribution of diagnosed colorectal cancers (not adenomas) at scale?	This is the linchpin of the "Yes" argument. If CADe only prevents future cancers, no 2028 impact is possible.	~60%
Can quasi-experimental designs (e.g., difference-in-differences) isolate AI's effect from secular trends by 2028?	Without exposure-tagged registries, attribution is speculative.	~50%
What is the true adoption rate of CADe in high-income countries, and how does it vary by healthcare system?	US data may not generalize to NHS, EU, or Japan.	~70%
Does the Corley ADR-mortality gradient hold at very high baseline ADRs (>40%)?	AI may push ADR beyond current benchmarks, where marginal benefit may diminish.	~50%

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
ADR (Adenoma Detection Rate)	The proportion of screening colonoscopies in which at least one adenoma is detected; a key quality metric in gastroenterology.
CADe (Computer-Aided Detection)	AI systems that assist endoscopists in detecting polyps during colonoscopy by highlighting suspicious areas in real time.
CISNET	The Cancer Intervention and Surveillance Modeling Network, a consortium of researchers who use simulation models to project the impact of cancer control interventions.
ctDNA	Circulating tumor DNA, fragments of DNA shed by tumors into the bloodstream, used as a biomarker for minimal residual disease (MRD) and treatment response.
DYNAMIC Trial	A clinical trial validating ctDNA-guided adjuvant therapy decisions in stage II colorectal cancer, published in NEJM 2022.
SEER	Surveillance, Epidemiology, and End Results program, a U.S. National Cancer Institute initiative that collects and publishes cancer incidence and survival data.
USPSTF	U.S. Preventive Services Task Force, an independent panel that makes evidence-based recommendations about clinical preventive services.

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.