Tribunal Session Summary

Session: conclave-20260513-144947-by-the-end-of-ad40

Depth: T6

Advocates: 5 | Judges: 6 | Cost: $4.7091 | Time: 37m 18s

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 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 impact on colorectal cancer survival rates in high-income countries by the end of 2028.

The strongest evidence for this verdict is the convergence of four interlocking constraints: the epidemiological dilution of AI's clinical benefits across national populations, the biological latency of the adenoma-carcinoma sequence, the 3–4 year reporting lag of cancer registries, and the absence of any credible, pre-specified mechanism for attributing observed survival changes to AI by 2028. While AI-assisted colonoscopy demonstrably improves adenoma detection rates (ADR) in trials and early-adopter clinics, the majority of colorectal cancers are diagnosed symptomatically—outside the reach of screening AI—and adoption will only reach 50–60% across high-income countries by 2028. Even granting optimistic effect sizes, this dilutes any population-level survival gain to less than 1 percentage point, below the threshold for materiality and indistinguishable from statistical noise and concurrent trends like FIT expansion and immunotherapy. The main caveat is that localized improvements in stage distribution and post-colonoscopy cancer rates may be detectable in high-adoption regions, but these do not constitute a material, cross-country survival impact.

Core Analysis

The Mechanism: AI's Pathway to Survival Improvement

The primary mechanism by which AI is expected to improve colorectal cancer (CRC) survival is through computer-aided detection (CADe) and diagnosis (CADx) systems used during screening colonoscopies. These systems, such as Medtronic's GI Genius and Olympus' Endo-AID, employ deep learning algorithms to analyze real-time endoscopic video feeds, highlighting potential polyps and adenomas that might be missed by human endoscopists.

The causal chain is well-established:

1. **AI → Higher Adenoma Detection Rate (ADR):** Randomized controlled trials (RCTs) and meta-analyses consistently show that CADe increases the ADR. A meta-analysis by Hassan et al. (Endoscopy, 2021) found a relative increase in adenomas detected per colonoscopy of approximately 44%. The Repici et al. (Gastroenterology, 2020) RCT demonstrated an absolute ADR increase from

40.4% to 54.8%.

2. **Higher ADR → Fewer Interval Cancers:** A higher ADR is a validated quality metric. The landmark Corley et al. study (N Engl J Med, 2014) established a strong observational link: each 1% absolute increase in ADR is associated with a 3% decrease in the risk of interval colorectal cancer (cancers diagnosed between scheduled screenings) and a 5% decrease in the risk of fatal interval cancer. This is known as the "Corley coefficient."
3. **Fewer Interval Cancers → Lower Mortality:** By preventing these interval cancers—many of which are aggressive and diagnosed at a late stage—AI-driven improvements in ADR are expected to reduce CRC mortality over time.

This mechanism is robust and forms the biological plausibility for AI's long-term impact. However, the critical question is not whether this chain exists, but whether it will produce a measurable, material effect on population-level survival rates by the end of 2028.

The Four Pillars of the "No" Argument

The majority verdict rests on four interlocking, evidence-based pillars that collectively make a material impact by 2028 impossible.

1. Epidemiological Dilution: The Denominator Problem

The most decisive argument against a 2028 impact is the mathematical dilution of AI's clinical benefits when scaled to a national population. Even if AI produces dramatic results in a controlled trial, its effect is diluted by three key factors:

- **Symptomatic Presentation:** Approximately 60% of colorectal cancers in high-income countries are diagnosed in patients who present with symptoms (e.g., bleeding, obstruction), not through routine screening. CADe systems only operate during screening or diagnostic colonoscopies for asymptomatic or FIT-positive individuals. They provide no benefit for the 60% of patients whose cancers are discovered too late for AI to have prevented them.
- **Partial Adoption:** Despite rapid progress, AI adoption will not be universal by 2028. Projections based on the Bass diffusion model and industry data suggest a weighted average adoption of 50–60% across high-income countries by 2028, with significant heterogeneity (e.g., ~60% in the US, ~75% in Japan, but only ~35% in the EU-27). This means only a fraction of screening colonoscopies will be AI-assisted.
- **Attributable Effect Size:** The effect size from the most optimistic real-world study, the Italian 2025 hybrid study (Lancet Gastro Hep), is a 10.2% absolute reduction in late-stage diagnoses within CADe-exposed, screen-detected cohorts.

Applying these factors to a national cohort of 100,000 annual CRC cases:

- Only 40,000 cases are screen-detected (60% symptomatic).
- Of those, only 20,000 are exposed to CADe (50% adoption).
- A 10.2% absolute reduction in late-stage diagnoses within this exposed cohort affects 2,040 patients.
- Shifting these 2,040 patients from a late-stage (e.g., Stage III) to an early-stage (e.g., Stage I) diagnosis yields a survival advantage. Using a conservative survival delta of 10 percentage points (Stage III to Stage II/I), the total survival gain is 204 life-years.
- When divided by the national denominator of 100,000 patients, this results in a **net population survival gain of +0.20 percentage points**.

Even under the most optimistic assumptions (e.g., a 44% survival delta from Stage IV to Stage I), the gain is only **+0.89 percentage points**. This is below the ≥1 percentage point threshold for "material" impact adopted by multiple advocates and is statistically indistinguishable from the background noise in national cancer registries.

2. Measurement and Reporting Lags

The gold standard for measuring cancer survival is the **5-year relative survival rate**. This metric requires five years of follow-up from the date of diagnosis to mature. For a patient diagnosed in 2028 to contribute to a 5-year survival statistic, they must be alive in 2033. Therefore, any survival benefit from AI used in 2025–2027 cannot be reflected in 5-year survival rates until the 2031–2035 reporting window.

While some registries, like the UK's NCRAS, offer rapid reporting of 1-year and 3-year survival data with less than a 12-month lag, these metrics are vulnerable to **lead-time bias**. If AI detects a tumor two years earlier than a human would have, the patient's "1-year survival from diagnosis" becomes nearly 100%, even if their actual lifespan is unchanged. This creates an artificial inflation of short-term survival rates that does not represent a true extension of life. The Tribunal concluded that without robust methods to adjust for this bias, 1-year survival data cannot be a reliable indicator of material impact.

3. The Attribution Problem

Even if a survival improvement were observed in 2028, isolating the contribution of AI from other concurrent, powerful trends would be impossible. High-income countries are simultaneously experiencing:

- **FIT Expansion:** National screening programs are expanding the age range for fecal immunochemical testing (e.g., UK extending to ages 50–54), which will independently shift stage distribution.
- **Post-COVID Catch-Up:** There is a massive backlog of delayed screenings and diagnoses from the pandemic, leading to a surge in early detections.
- **Therapeutic Advances:** Immunotherapy for MSI-high tumors (e.g., pembrolizumab) has significantly improved survival for a subset of patients.

DeepSeek R1 claimed that a pre-specified difference-in-differences (DiD) analysis in the UK’s NBOCA 2028 report would provide causal attribution. However, this claim was deemed **unverifiable and speculative** by the judges. No such protocol has been publicly pre-registered, and the timeline for such an analysis to be published by December 31, 2028, is implausible. Without a credible, pre-registered method to disentangle AI’s effect, any observed change cannot be attributed to it.

4. Adoption and Integration Timelines

While adoption is faster than initially projected, it is not fast enough. The Bass diffusion model, with an innovation coefficient (p) of 0.035–0.05 and an imitation coefficient (q) of 0.40–0.45, projects a procedure-weighted exposure of 50–55% by 2028. This is insufficient for a population-level signal. Furthermore, integration challenges—such as clinician training, workflow adaptation, and reimbursement variability (e.g., absent in many EU countries until 2026)—act as bottlenecks that prevent the rapid, uniform scaling required for a measurable impact.

Main Potential Impact Areas for AI

While a material survival impact is not expected by 2028, AI is poised to make significant contributions in several areas:

- **Screening & Early Detection (CADe/CADx):** This is the most mature application. AI acts as a "second set of eyes" during colonoscopy, reducing polyp miss rates and improving characterization. The primary benefit will be a reduction in post-colonoscopy colorectal cancer (PCCRC) rates and a shift in stage distribution among screen-detected cancers.
- **Digital Pathology:** AI can predict microsatellite instability (MSI) directly from H&E-stained slides, potentially accelerating access to immunotherapy. It can also standardize the assessment of tumor budding and the tumor-stroma ratio, which may refine adjuvant therapy decisions.
- **Risk Stratification:** AI models can analyze electronic health records (EHR) to identify high-risk patients who are non-adherent to screening, enabling targeted outreach and increasing screening uptake.
- **Operational Efficiency:** AI can optimize scheduling, reduce time-to-colonoscopy after a positive FIT, and standardize surveillance intervals, improving the overall efficiency of CRC care pathways.

Expected Timeframe and Why

A material, population-level impact on colorectal cancer survival rates in high-income countries is expected between **2031 and 2035**.

This timeframe is determined by the convergence of three factors:

1. **Adoption Saturation:** AI-assisted colonoscopy is projected to reach 70–80% of screening and diagnostic procedures across high-income systems by 2028–2030.
2. **Epidemiological Latency:** The adenoma-carcinoma sequence takes 5–10 years. Cancers prevented by AI-detected and removed adenomas in the late 2020s will begin to manifest as a reduction in diagnosed cancers around 2029–2031.
3. **Measurement Maturation:** The 5-year survival statistics for cohorts first exposed to widely adopted AI (circa 2025–2027) will be reported in the 2031–2035 publication cycle. By this time, attribution studies will also have matured, allowing for a credible assessment of AI’s contribution.

How We Got Here

Scorecard

Judge	GPT-5	Claude Opus 4.7	Gemini 3.1 Pro	Grok 4	DeepSeek R1	Winner
MiniMax M2.7	ACCEPT	REJECT	REJECT	REJECT	REJECT	GPT-5
Qwen 3 235B (Cerebras)	REJECT	REJECT	REJECT	REJECT	REJECT	(REMAND)
Qwen 3.5 397B	REJECT	REJECT	REJECT	REJECT	REJECT	(REJECT)

Judge	GPT-5	Claude Opus 4.7	Gemini 3.1 Pro	Grok 4	DeepSeek R1	Winner
DeepCogito v2.1 671B	ACCEPT	REJECT	REJECT	REJECT	REJECT	GPT-5
GPT-OSS 120B	ACCEPT	REJECT	REJECT	REJECT	REJECT	GPT-5

Convergence Assessment

The convergence was **epistemic**, not affective. The four NO advocates (GPT-5, Claude Opus 4.7, Gemini 3.1 Pro, Grok 4) maintained distinct mechanisms—adoption lag, attribution, cohort dilution, and biological latency—but converged on the same conclusion through evidence-driven reasoning. Gemini 3.1 Pro and Claude Opus 4.7dvocate-A independently derived the same dilution arithmetic (~0.5-1.0 pp gain) without coordinating, demonstrating that the math is robust. DeepSeek R1 showed medium sycophancy risk, as it maintained its YES conclusion despite withdrawing multiple claims and shifting its primary endpoint, suggesting a desire to preserve a positive stance rather than follow the evidence.

Closing Statements

GPT-5: Maintained a "No" position, refining its argument to acknowledge faster adoption but emphasizing that a procedure-weighted 50–55% exposure by 2028 is insufficient for a material, cross-country survival impact. The strongest evidence was the ADR → interval cancer → mortality causal chain, supported by robust RCTs and the Corley coefficient, but constrained by measurement lags that push observable effects to 2031–2035.

Claude Opus 4.7: Held a "No" position, defining materiality as a ≥1 pp improvement in 5-year survival with credible attribution. The strongest evidence was the cohort-dilution arithmetic, which showed that even the Italian study's 35% stage-shift dilutes to a 0.5–1.0 pp population gain. The binding constraint is attribution, as no registry will publish a causal analysis by 2028.

Gemini 3.1 Pro: Maintained a "No" position, conceding that AI can produce immediate stage-shifts but arguing that the effect is mathematically too small at the population level. The strongest evidence was the "denominator dilution" model, which calculated a net survival gain of only +0.20 to +0.89 pp, and the inescapability of lead-time bias in 1-year survival metrics.

DeepSeek R1: Maintained a "YES" position after significant refinement, shifting from survival rates to stage-shift and PCCRC reduction as the primary impact metrics. The strongest evidence was the Italian 2025 study showing a 35% drop in late-stage diagnoses and the claim of a pre-specified DiD analysis in the NBOCA 2028 report, though the latter was deemed unverifiable.

Reality Checks

Claim	Challenge	Result
DeepSeek R1: "AI will avert 12,000+ colorectal deaths annually in high-income nations by 2028"	**Claude Opus 4.7:** "Show the arithmetic. The Corley translation requires 5–10 years for mortality reduction. How do adenomas removed in 2025–2027 prevent 12,000 deaths in 2028?"	**Conceded and withdrawn.** DeepSeek R1 admitted the claim was overstated for 2028 and redefined impact as stage-shift, not mortality reduction.
DeepSeek R1: "Deep learning models achieve 98% accuracy in detecting colorectal cancer from biopsy slides (Bándi et al., Nature Medicine 2019)"	**Claude Opus 4.7:** "That arXiv ID does not correspond to a real NHS screening study to my knowledge. Provide the actual citation or withdraw it. This looks like a hallucinated reference."	**Revised and corrected.** DeepSeek R1 conceded the error, citing Kather et al. (Lancet Digit Health 2023) with 92.4% accuracy instead.
DeepSeek R1: "Paige Prostate is FDA-cleared for prostate cancer; colorectal adaptations are in late-stage trials (Paige FDA submission Q4 2023)"	**Claude Opus 4.7:** "Provide the submission docket number or retract."	**Revised.** DeepSeek R1 corrected the timeline to Q1 2024 (docket DEN240001) and acknowledged the product was pending approval.

Next Steps

Next Step	Why It Matters	Confidence
Monitor the publication of the UK's NBOCA 2028 report for a pre-specified difference-in-differences analysis on AI's impact.	This is the linchpin of the YES argument. If such an analysis is published and shows a significant, attributable effect, it would challenge the Tribunal's verdict.	20%
Track real-world adoption rates of CADe systems in the US, EU, and Japan through 2028.	The Bass diffusion model projects 50-60% weighted average adoption. If adoption significantly exceeds this (e.g., >70%), it could compress the timeline for a population-level signal.	60%
Assess the true effect size of CADe on advanced adenoma and early cancer detection in routine practice, not just trials.	The Italian study reports a 35% relative reduction in late-stage diagnoses, but this may be inflated by selection bias. A lower real-world effect size would further reduce the projected population impact.	70%
Evaluate how national cancer registries plan to adjust for lead-time bias in short-term survival metrics.	Without adjustment, reported survival improvements from AI may be misleading. This is critical for interpreting any early signals.	60%

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 for endoscopist performance.
CADe (Computer-Aided Detection)	AI systems that analyze real-time endoscopic video to highlight potential polyps or adenomas during a colonoscopy.
CADx (Computer-Aided Diagnosis)	AI systems that analyze endoscopic images to predict the optical histology (e.g., hyperplastic vs. adenomatous) of a detected polyp.
Corley coefficient	An observational finding that each 1% absolute increase in adenoma detection rate (ADR) is associated with a 3% reduction in the risk of interval colorectal cancer.
FIT (Fecal Immunochemical Test)	A non-invasive screening test that detects hidden blood in the stool, used to identify individuals at higher risk for colorectal cancer.
Interval colorectal cancer	A colorectal cancer diagnosed between scheduled screening colonoscopies, often indicating a missed lesion during the prior procedure.
PCCRC (Post-Colonoscopy Colorectal Cancer)	A subset of interval cancers, specifically those diagnosed within a defined period (e.g., 3 or 5 years) after a colonoscopy.
SEER (Surveillance, Epidemiology, and End Results)	A program of the National Cancer Institute that collects and publishes cancer incidence and survival data from population-based cancer registries in the United States.
Stage-shift	A change in the distribution of cancer diagnoses toward earlier stages (e.g., more Stage I, fewer Stage IV), which is associated with improved survival outcomes.

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 (e