ChatGPT 100,000 Patient 24-Month *In Silico* Phase III 5-Arm Pancreatic Cancer Clinical Trial Triplicate

Kevin Kawchak

Chief Executive Officer

ChemicalQDevice

San Diego, CA

July 24, 2025

kevink@chemicalqdevice.com

Abstract

Inquiry: Is it possible for ChatGPT to simulate three reproducible 100,000 patient pancreatic ductal adenocarcinoma (PDAC) Phase III clinical trial reports? If so, can the results be internally and externally validated, cross-verified using other AI models, and be compared both clinically and financially to other trials?

Concept: 5 arms based on the Daraxonrasib + Mitazalimab + liposomal Irinotecan drug combination, baseline characteristics, and patient archetypes were identified from a prior study: doi.org/10.5281/zenodo.15735068. Six artificial intelligence models were then implemented to address the clinical trial pipeline: o3ph: ChatGPT o3-pro Research, g25p: Google Gemini 2.5 Pro, grk4: Grok 4, grk3: Grok 3 Think, o3pr: ChatGPT o3-pro, and ops4: Opus 4 Extended. o3ph generated the ICH E3-aligned trial reports, log files, plus internal, and external validations. g25p, grk4, grk3, o3pr, and ops 4 provided cross verifications that highlighted trial-to-trial and model-to-model correlations. g25p utilized 24 generations in the study to produce a virtual trials overview, while o3ph provided a meta-analysis of pooled and scored data versus relevant virtual and on-site trials. o3ph also provided a financial assessment and value proposition of USD estimates against Phase II and Phase III studies. ops4 provided visualizations written in Python for the majority of the sections.

Results: 100,000 individual patients generated from three separate o3ph conversations followed multiplicative hazard ratios and per-arm monthly hazards set in the prompt. Key variables were independent of each other, which yielded distributions of uncensored results. Log file cumulative effects of the censored 100,000 patients yielded expected results in OS by Arm (A > D > E), ≥G3 AE (A > D > E), and PFS (A > D > E). Baseline characteristics by metric across trials were in close alignment, and internal validations between log files and trial reports exhibited similar performance. External validation vs. a Flatiron Health dataset for OS passed, while ECOG validation saw higher differences. These deviations, along with a KRAS-mutant labeling issue were high, but consistent in magnitude across the three trials.

Outputs: In order to consolidate trial information, validations, and cross-verifications, g25p processed 24 of these relevant outputs to create a virtual trials overview. The core trial information, technical specifications, reproducibility, and validation findings provided a concise output needed for subsequent comparisons to trials. The method used to pool the current study with prior studies was accomplished by o3ph utilizing the virtual trials overview alongside online clinical trial data to produce a 9,574 word meta-analysis. Results focused on PRODIGE-4 and NAPOLI-1 trials that were top two in OS, while Arm A was third. However, the Arm D doublet of Daraxonrasib + Mitazalimab was less toxic than the other trials, and was found to be more clinically feasible than FOLFIRINOX in PRODIGE-4.

Impacts: The financial assessment and value proposition performed by o3ph and visualized by ops4 placed an estimated price of $36,330 on the current study (1 user at $150/hr working 60 hrs/wk). Estimates for other virtual trials ranged from $120,000-$600,000, while a real Phase II trial was $20.0M, and the Phase III trial estimate was $100.0M. Time-to-decision was fastest for the 100K Triplicate at 1 month, while other studies ranged from 4.5 months to 5.0 years. The AI's main financial decision was that Arm A (Daraxonrasib + Mitazalimab + liposomal Irinotecan) was not a strong enough candidate, and the results from the current study were estimated to save $19.96M to avoid a clinical trial failure. In addition, a $2.36M burn rate reduction was anticipated, with an overall cost reduction of 99.9997% vs. a Phase III trial per patient.

Outcome: The main benefit was that reproducibility was observed across a single trial or multiple trials, while individual patients likely varied based on raw exponential sampling. The o3ph feat was primarily in providing a trial report that was replicable between the other trials performed in separate conversations. Similarly, the g25p model's processing of 24 outputs to create a virtual trials overview could not me accomplished by any other model due to token limitations. The overview served to inform the final meta-analysis and financial assessment by o3ph, providing tangible comparisons and planning tools for upcoming studies. All work was performed by one user in a 30 day window.

Table of Contents

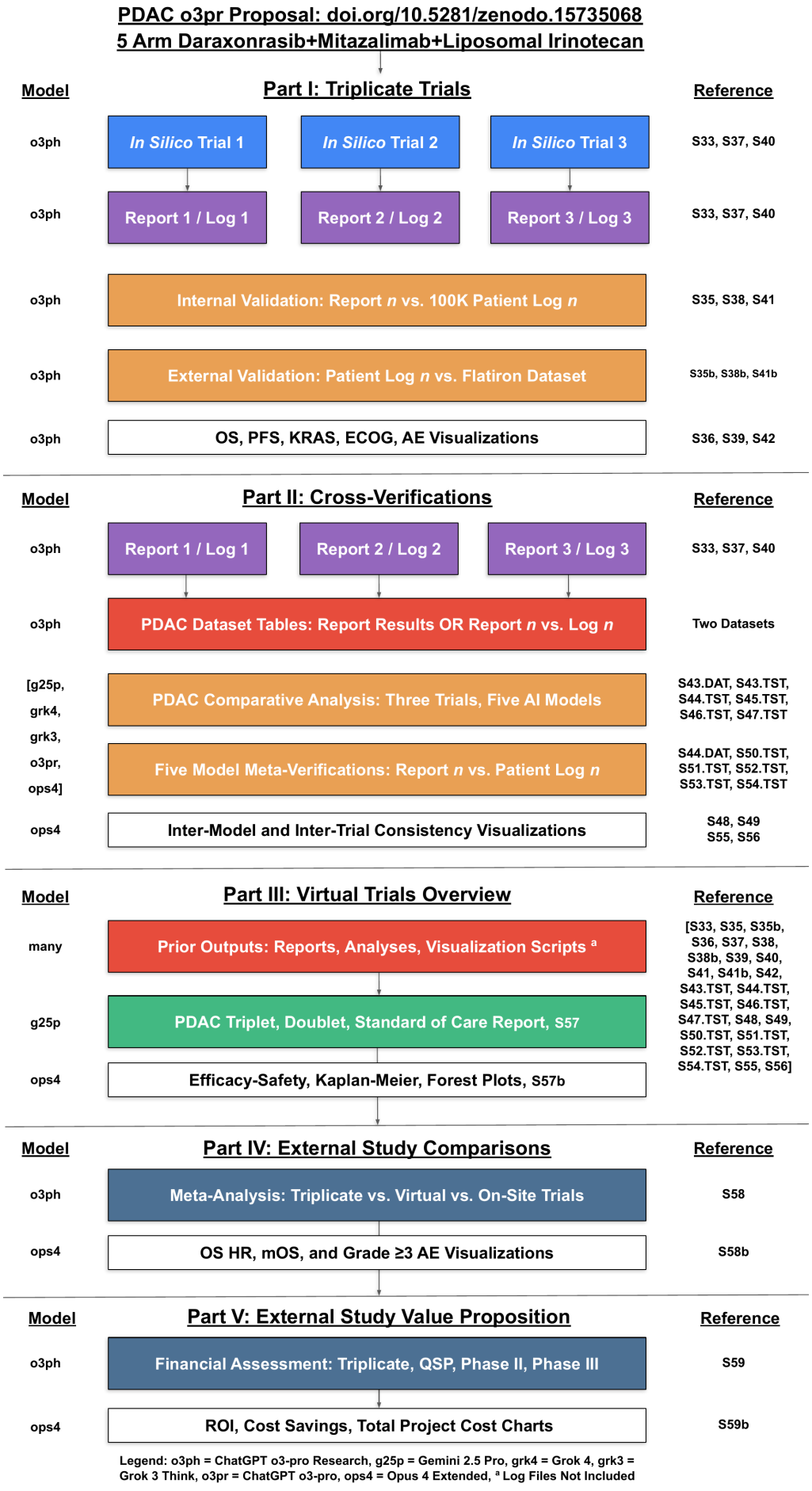

Figure 1: PDAC 100K Patient In Silico Clinical Trial Pipeline

1 Introduction

1.1 LLMs Benefit In Silico Studies

In 2025, clinical trial based datasets have been produced by combining oncology high context conversational AI reports and meta-analyses [1, 2, 3]. Output lengths of 10,000 words or more consistently exceeded the lengths of full-length articles, with AI composing multiple relevant sections in less than an hour. AI informed in silico clinical trial proposals containing patient, drug, and financial information were generated using several readily available models, including o3pr, ops4, g25p, and o3ch [1]. Proposal verification, validation & uncertainty quantification plans, as well as trial protocol and no-go criteria were included using AI by Kawchak K. Visualizations in Python yielding multiple chart types were best generated using ops4 or son4, although publication ready images required additional screening. Clinical trial data represented by risk of bias, forest plots, heatmaps, budgets, and financial timelines have been generated effectively. Meta-prompting using an AI model to generate, refine, or interpret prompts has been an effective tool to improve the output quality of subsequent generations. This method was particularly effective when large datasets needed to be implemented, but were not fully understood by the user.

For instance, meta-prompts were used with g25p on a 408,081 word dataset to determine a more optimized prompt for processing the large dataset. The g25p 1M token context length has enabled workflows that other leading AI models could not achieve. Verifications using the same prompt across multiple separate models was another effective technique to gain a consensus by Kawchak K. Five models were utilized to judge five virtual clinical trial proposals in best of 10 scoring, to yield a consensus on PDAC digital twin proposals [1]. Criteria were based on deliverable completion, citation ability, trial impact, and funding probability, with the overall score being highest for o3pr at 9.09. Additional analysis of the three top proposals yielded trial timelines, a 6-part ROI analysis, FTE allocation, and budget comparisons. For small tasks, models such as 4och [4, 5, 6] and grk3 [1, 2] were effectively utilized for fast insights regarding well known information.

1.2 In Silico Studies, Local Trials

In early 2024, Arcus Biosciences sponsored the PRISM-1 study, with Dr. Zev Wainberg, stating that "There is already a lot of data from randomized phase II and III trials on patients treated with gemcitabine/nab-paclitaxel. A synthetic arm is really very efficient since it reduces the number of patients needed for a study so the trial timeline is shorter and costs are reduced [7]." Later in 2024, Arcus Biosciences utilized data from a prior Phase 1B ARC-8 study which yielded a "37% reduction in risk of death and a 5.9-month improvement in median overall survival" for patients treated with quemliclustat-based regimens when compared to a 122 patient Synthetic Control Arm® of patients treated with chemotherapy alone in a post-hoc analysis. [8].

A 2024 *Nature Cancer* article by researchers at Johns Hopkins and Cedars-Sinai Medical Center featured a Molecular Twin AI platform that integrated a dataset of 6,363 clinical and multi-omic molecular features to predict outcomes for pancreatic adenocarcinoma patients. "Our platform enables discovery of parsimonious biomarker panels and performance assessment of outcome prediction models learning from resource-intensive panels. This approach has considerable potential to impact clinical care and democratize precision cancer medicine worldwide [9]."

In 2024, Asghar et al. published on the "Prediction of therapeutic response and cancer outcomes in solid tumours via in silico clinical trials." "For all 8 clinical studies, the digital twin model accurately simulated both trial arms, compared drug efficacy across arms and predicted which treatment was most active. "Blinded evaluation: Using data for paclitaxel, the model correctly predicted that nab-paclitaxel+gemcitabine response rates were higher than gemcitabine in metastatic pancreatic cancer (predicted LOR -0.090, p = <0.001) [10]." Toshimoto et al. in November 2024 published on an immune-oncology quantitative systems pharmacology (IO-QSP) model with custom tumor diameter, growth rate and immune cell proportions parameters. Two mechanisms of action were considered, with the authors "successfully reproducing the clinical responses of anti-PD-1 and/or combination therapy with anti-PD-1 and anti-CTLA-4". "The established IO-QSP models captured clinical responses of standard of care treatments and checkpoint inhibitors in both gastric and pancreatic cancers [11]."

The 2025 Phase III AVATAR Trial by Sarno et al. for personalized pancreatic cancer treatment featured whole-exome sequencing + mouse/ PDO drug screens run. Avatar data selected matched therapy candidates; although only 10% received matched drugs, that subgroup doubled median OS (19.3 mo vs 8.7 mo). "The study showed that personalized medicine did not improve survival as compared with standard of care in an intention-to-treat population [12]." The Ko et al. April 2025 article on "Investigational Use of Real-World Data as a Hybrid Control in Pancreatic Ductal Adenocarcinoma From the Randomized Phase Ib/II MORPHEUS Trial" utilized a hybrid control arm and an experimental arm (atezolizumab + PEGylated recombinant human hyaluronidase (Atezo + PEGPH20)). HRs ranged from 1.02 to 1.06, and were comparable with the reported trial HR (HR 0.91; 95% CI: 0.56, 1.49), with precision improvements experienced when using the hybrid control [13]." Pourmousa et al. in April 2025 detailed an AI methodology by researchers at UNC and MIT for pancreatic cancer that "evaluates predictive approaches to identify synergistic drug combinations using a dataset from the National Center for Advancing Translational Sciences (NCATS)." "Screening 496 combinations of 32 anticancer compounds against the PANC-1 cells experimentally determined the degree of synergism and antagonism." "Beyond highlighting the potential of ML, this work delivers 307 experimentally validated synergistic combinations, demonstrating its practical impact in treating pancreatic cancer." [14]."

2 Methods

2.1 In Silico Trials, Logs, Validations

The five arm drug combinations were based on a top TSVS score (8.15) of Daraxonrasib + Mitazalimab + liposomal Irinotecan, along with baseline characteristics, archetypes, and timeline from Proposal A in a recent Kawchak K. article: doi.org/10.5281/zenodo.15735068 [1]. Weibull shape parameter, multiplicative hazard ratios, synergy factors, biomarker adjustments, and event times were further implemented to assist ICH E3-Compliant report results and validations. 100,000 single row synthetic records were utilized for the virtual patients across eleven clinical indicators. 100,000 patient entries for each trial were obtained by requesting a copy of the log in a follow up prompt in the same conversation from o3ph. Internal validations were conducted with Prompt 32 that identified discrepancies between three of the trial tables and respective patient log files (S35, S38, S41). The term 'triplicate' refers to the three trial study, while the term 'triplet' refers to Arm A, and 'doublet' can refer to Arms B, C, or D; with Arm E being the control.

Quantitative data from external validation Prompts 34-36 utilized the Flatiron Health abstract which was further extrapolated by o3ch into usable Prompts (S35b, S38b, S41b) [15]. The prompts were then run against each of the patient log files separately to yield three sets of validations results. Thirty chart visualizations based on uploading log files with Prompt 30 into a new conversation were obtained with o3ph. Charts were then screened by the user for potential article incorporation. Cross-verifications were performed between three trial sets of five tables (S43.DAT.02.TAB) across five AI models (g25p, grk4, grk3, o3pr, ops4) to reveal inter-model and inter-trial consistencies (S43.TST, S44.TST, S45.TST, S46.TST, S47.TST). The results from each of the models were then visualized in charts by ops4. A second set of cross-verifications were performed based on meta-verifications of report tables vs. patient log files between the three virtual trials; followed by ops4 visualizations indicating cross-model and trial-to-trial relationships. The Prompt 43 trial overview was derived from meta-prompts utilizing Proposal A from a prior article [1], along with other generations in this study shown. The virtual trial overview by g25 (S57.REP.01.P43) consisted of 24 generations in the study as inputs; including reports, analyses, and visualization scripts to gain a better understanding of the pipeline. The trial overview included benefits, drawbacks, and reproducibility findings; and referenced input generations. Clinical trial related charts such as Kaplan-Meier and Forest plots were then obtained by ops4 (S57b.VIS.01.P43b), and included in the supporting files. The prompts in the article were written in LaTeX, see supporting files for formatting used in study.

2.2 External Studies: Meta-Analysis, Economics

External study comparisons versus known virtual and on-site trials were conducted by o3ph (S58.REP.02.P44) using Prompt 44 with the S57.REP.01.P43 trial overview as an input, and Deep research to obtain a 9,574 word in-depth meta-analysis regarding best performing study arms against two in silico trials, as well as three Phase III trials: MPACT, NAPOLI-1, and PRODIGE 4. Additional pooled clinical metrics and head-to-head efficacy–toxicity scoring of experimental and control arms of the current study versus the three Phase III trial outcomes were defined in the prompt. The prompt also provided recommendations to address research gaps for the current study. Subsequent visualizations focusing on head-to-head trial analyses were obtained in two sets of ten charts by ops4.

The financial assessment by o3ph utilized the S58.REP.02.P44 meta-analysis and S57.REP.01.P43 trial overview as inputs to provide a value proposition of the current vs. external single run virtual trials, and a mechanistic QSP model with more patient data granularity. The prompt specified details regarding labor cost, cloud compute, and project duration. Comparisons regarding in-person Phase II and Phase III trial estimates for budget, timeline, and cost per patient were also included in the prompt. Additionally, a grant funding justification framework was included that specified study outcomes with references of optimizations and methodologies that added value. Charts focusing on financial estimate by ops4 were run twice to yield a larger pool of publishable images. Content from the Introduction was obtained using references generated by o3ch and grk4, with the user composing the section. The author of the article conducted experiments, analyses, and composed the article.

AI software were based on unmodified chat-based inference LLMs with generation times obtained using a digital stop watch. The following models (current study in bold) were accessed through MacOS 14.5 (23F79) and Google Chrome browser Version 137.0.7151.120. ChatGPT Deep research is also referred to as ChatGPT Research. See prior methods for o3re, son4, 45dr.

AI Models:

1. **o3ph:** OpenAI ChatGPT o3-pro Deep research was accessed through the ChatGPT website pro chat interface [16, 17]
2. **g25p:** Google Gemini 2.5 was accessed through the Google AI Studio playground. Settings: Temp=1, Thinking mode=On, Set thinking budget=32768, All Off=(Structured output, Code execution, Function calling, Grounding with Google Search, URL context), All Off=Safety settings, Output length=65536, Top P=0.95. [18, 19]
3. **grk4:** xAI Grok 4 was accessed through the Grok SuperGrok website chat interface [20]
4. **grk3:** xAI Grok 3 Think was accessed through the Grok SuperGrok website chat interface [21]
5. **o3pr:** OpenAI ChatGPT o3-pro was accessed through the ChatGPT website pro subscription chat interface [22, 23]
6. **ops4:** Claude Opus 4 Extended was accessed through the Claude website professional plan chat interface [24, 25]
7. o3ch: OpenAI ChatGPT o3 was accessed through the ChatGPT website pro subscription chat interface [16, 17]
8. 4och: ChatGPT 4o was accessed through the ChatGPT website pro, mobile teams (iOS) for supplementary research [26]
9. o3re: OpenAI ChatGPT o3 Research was accessed through the ChatGPT website pro subscription chat interface [1, 16]
10. son4: Claude Sonnet 4 Extended was accessed through the Claude website professional plan chat interface [1, 27]
11. 45dr: OpenAI ChatGPT 4.5 Deep research was accessed through the ChatGPT website pro subscription interface [3, 28].

Patient Triplicate Log File Analysis

Trial 1

patient_id	arm	archetype	age	stage_iv	ecog	kras_g12c	gbrca	ca19_9	time_to_progression_or_death	time_to_death	time_to_first_G3_AE
000001	E	ARCH-01	56.0	1	0	0	0	3981.9	8.45	23.41	26.13
023649	D	ARCH-02	73.6	1	1	0	0	7648.3	3.80	31.29	1.05
046253	C	ARCH-03	61.2	1	1	0	0	1142.8	14.84	1.86	23.85
050205	C	ARCH-04	66.3	1	0	0	1	2922.8	0.91	20.13	0.51
057254	A	ARCH-05	84.0	1	1	1	0	299.8	2.36	8.87	3.67
069416	D	ARCH-06	67.8	1	0	0	0	4933.9	4.25	4.74	4.24
100000	B	ARCH-07	65.1	1	1	0	0	9682.2	10.14	1.26	16.61

Patient 000001 had stage IV disease (ECOG 0), KRAS G12C -, BRCA wild-type, CA19-9 3981.9 U/mL; OS 23.4 mo, PFS 8.5 mo.
Patient 023649 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 7648.3 U/mL; OS 31.3 mo, PFS 3.8 mo.
Patient 046253 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 1142.8 U/mL; OS 1.9 mo, PFS 14.8 mo.
Patient 050205 had stage IV disease (ECOG 0), KRAS G12C -, BRCA mutated, CA19-9 2922.8 U/mL; OS 20.1 mo, PFS 0.9 mo.
Patient 057254 had stage IV disease (ECOG 1), KRAS G12C +, BRCA wild-type, CA19-9 299.8 U/mL; OS 8.9 mo, PFS 2.4 mo.
Patient 069416 had stage IV disease (ECOG 0), KRAS G12C -, BRCA wild-type, CA19-9 4933.9 U/mL; OS 4.7 mo, PFS 4.3 mo.
Patient 100000 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 9682.2 U/mL; OS 1.3 mo, PFS 10.1 mo.

Trial 2

patient_id	arm	archetype	age	stage_iv	ecog	kras_g12c	gbrca	ca19_9	time_to_progression_or_death	time_to_death	time_to_first_G3_AE
000001	B	ARCH-01	68.3	1	0	0	0	4013.9	16.11	20.21	10.38
023649	E	ARCH-02	79.3	1	1	0	0	2055.6	2.34	3.26	10.17
046253	A	ARCH-03	46.4	0	1	0	0	1558.7	2.38	0.37	16.06
050205	A	ARCH-04	67.0	1	0	0	1	1516.4	2.97	8.11	0.97
057254	D	ARCH-05	62.4	1	0	1	0	5598.9	1.48	1.85	24.00
069416	A	ARCH-06	68.9	1	1	0	0	352.1	1.04	11.20	2.07
100000	C	ARCH-07	58.1	1	1	0	0	5540.9	2.96	1.00	2.59

Patient 000001 had stage IV disease (ECOG 0), KRAS G12C -, BRCA wild-type, CA19-9 4013.9 U/mL; OS 20.2 mo, PFS 16.1 mo.
Patient 023649 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 2055.6 U/mL; OS 3.3 mo, PFS 2.3 mo.
Patient 046253 had non-metastatic disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 1558.7 U/mL; OS 0.4 mo, PFS 2.4 mo.
Patient 050205 had stage IV disease (ECOG 0), KRAS G12C -, BRCA mutated, CA19-9 1516.4 U/mL; OS 8.1 mo, PFS 3.0 mo.
Patient 057254 had stage IV disease (ECOG 0), KRAS G12C +, BRCA wild-type, CA19-9 5598.9 U/mL; OS 1.9 mo, PFS 1.5 mo.
Patient 069416 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 352.1 U/mL; OS 11.2 mo, PFS 1.0 mo.
Patient 100000 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 5540.9 U/mL; OS 1.0 mo, PFS 3.0 mo.

Trial 3

patient_id	arm	archetype	age	stage_iv	ecog	kras_g12c	gbrca	ca19_9	time_to_progression_or_death	time_to_death	time_to_first_G3_AE
000001	B	ARCH-01	68.3	1	0	0	0	4013.9	16.96	7.75	5.51
023649	A	ARCH-02	75.3	1	1	0	0	14695.1	14.00	24.00	19.16
046253	E	ARCH-03	65.7	0	1	0	0	10.0	1.16	2.09	13.18
050205	E	ARCH-04	51.1	1	1	0	1	4548.5	1.14	3.08	17.51
057254	A	ARCH-05	48.1	1	1	1	0	2080.0	17.96	13.87	4.21
069416	B	ARCH-06	54.9	1	1	0	0	7455.2	9.43	12.88	1.93
100000	C	ARCH-07	60.1	1	1	0	0	14251.4	14.39	2.45	2.00

Patient 000001 had stage IV disease (ECOG 0), KRAS G12C -, BRCA wild-type, CA19-9 4013.9 U/mL; OS 7.8 mo, PFS 17.0 mo.
Patient 023649 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 14695.1 U/mL; OS 24.0 mo, PFS 14.0 mo.
Patient 046253 had non-stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 10.0 U/mL; OS 2.1 mo, PFS 1.2 mo.
Patient 050205 had stage IV disease (ECOG 1), KRAS G12C -, BRCA mutated, CA19-9 4548.5 U/mL; OS 3.1 mo, PFS 1.1 mo.
Patient 057254 had stage IV disease (ECOG 1), KRAS G12C +, BRCA wild-type, CA19-9 2080.0 U/mL; OS 13.9 mo, PFS 18.0 mo.
Patient 069416 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 7455.2 U/mL; OS 12.9 mo, PFS 9.4 mo.
Patient 100000 had stage IV disease (ECOG 1), KRAS G12C -, BRCA wild-type, CA19-9 14251.4 U/mL; OS 2.5 mo, PFS 14.4 mo.

Table 1: Seven Patient/Archetype Analyses, o3ph. Ref: S33, S37, S40

3 Part I: PDAC Triplicate Trials

3.1 Patient Triplicate Log File Analysis

In order to evaluate clinical metrics of generated trials derived from Prompt 30 using o3ph, seven patient's data corresponding to each of the seven archetypes was assisted by o3pr and o3ch. Table 1 illustrates the 12 variables including patient age, kras_g12c status, progression-free survival, and overall survival for each of the patients. Patients 000001 and 100000 were held constant, and represented ARCH-01 and ARCH-07. Patients corresponding to ARCH-02 through ARCH-06 were randomly selected by o3pr. Trials 2 and 3 utilized the same patient ids as with Trial 1. For Trial 1 Patient 000001, 26.13 months exceeded the trial timeline, but for the raw patient-level dataset values were likely untouched exponential draws. This means that uncensored values in the raw data are likely acceptable.

Overall, many of the patient data followed expected mathematical distributions such as Trial 1 "Patient 000001 had stage IV disease (ECOG 0), KRAS G12C -, BRCA wild-type, CA19-9 3981.9 U/mL; OS 23.4 mo, PFS 8.5 mo." However, values such as Trial 1 Patient 046253 time to first ≥G3 AE (23.85 mo) was greater than OS (1.86 mo). The likely reason for these events, which were likely uncensored compared to the trial reports stem from Prompt 30's instruction: "Use Kaplan-Meier analysis on time_to_progression_or_death (for PFS) and time_to_death (for OS), censored at 24 months." Cumulative results such as in Table 6-1 of the three trials reflect expected ranges of mPFS ~4 mo vs. mOS ~8 mo for Arm A. External validations also showed OS consistencies across trials vs. the Flatiron Health dataset.

PDAC Triplicate Trials Patient Log Files, o3ph

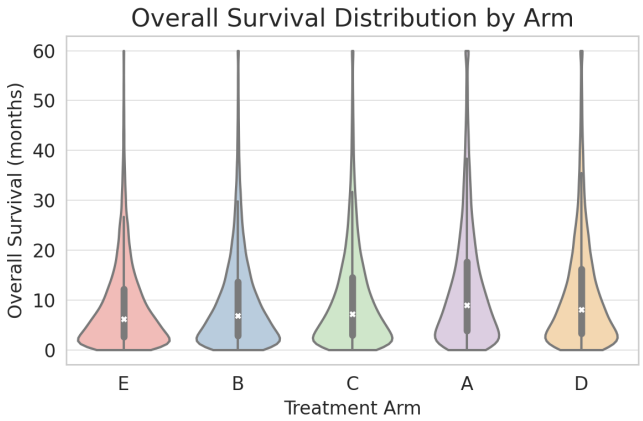

Figure 2: Trial 1 Violin Plot of Treatment Arms

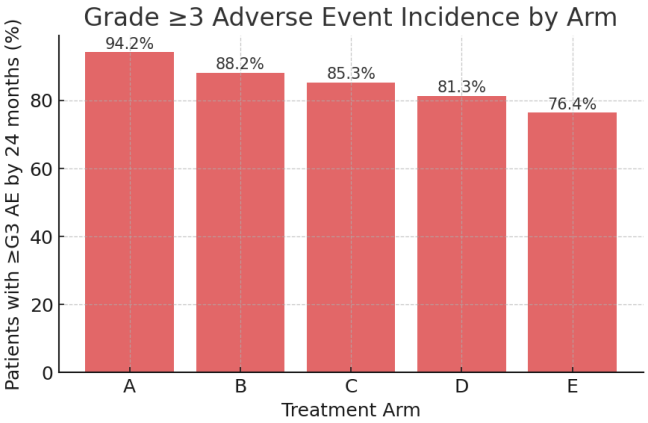

Figure 3: Trial 2 Grade ≥3 AE Incidence by Arm

E: Control Arm, A: Main Experimental
Trial Arms OS: A > D > C > B > E

Lower # of Adverse Events is Better
Grade ≥3 AEs: E < D < C < B < A

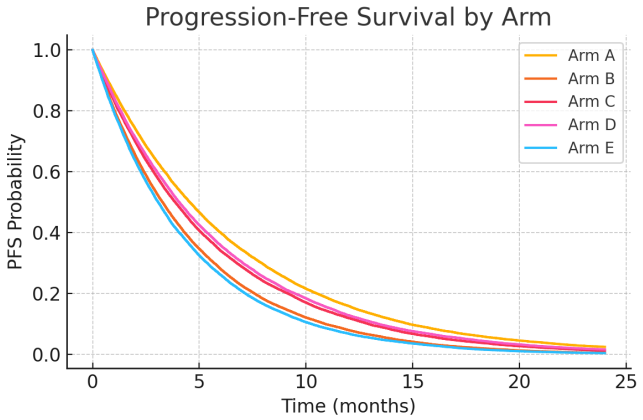

Figure 4: Trial 3 Kaplan-Meier PFS Curves by Arm

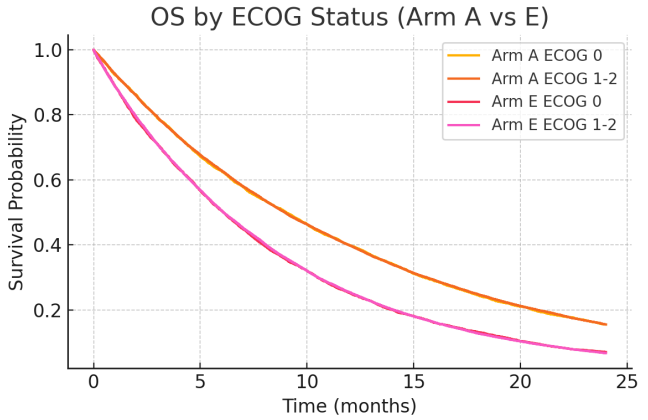

Figure 5: Trial 3 OS by ECOG Status

Higher Kaplan-Meier Curve is Better
Trial Arms PFS: A > D > C > B > E

Highest Survival for Arm A ECOG 0/1/2
Patient Lifespan of Arm A > Control

Each of the log files were re-analyzed by o3ph to yield visualizations. Figure 2 shown above represents the 100,000 patient monthly OS distribution across the five arms. Based on Prompt 30 Section 2.2 OS multiplicative HRs of lower OS HRs Daraxonrasib vs. baseline of 0.85 vs. Mitazalimab 0.90 (lower is better), combined with the 0.90 synergy factor for the triplet (Arm A): the shown best performance of Arm A and worst performance for Arm E is rational. Figure 3 also correlates with Prompt 30, as Arm A monthly probability for a Grade ≥3 AE was highest at 0.12 and lowest for the control at 0.06. The results for PFS by Arm in Figure 4 also correspond to Prompt 30 favorable multiplicative HR PFS assignments for Daraxonrasib at 0.80 vs. Mitazalimab at 0.95.

3.2 Internal and External Validations

The following page illustrates Table 5-1 trial report consistencies for the 100,000 patients across age, stage, ecog status, KRAS-mutant, etc. Many values showed favorable differences to the tenth of a value. The following internal validation regarding report values vs. patient log values in Table 2 also showed minor deviations across most values. The exception was the KRAS-mutant (%), with deviations between report values and log values of 85.8%, 86.1%, 86.2%. The large differences are believed to primarily stem from the reports including KRAS-mutant (>90%) in pancreatic cancer vs. the log file specifying Kras G12C (<1–2% of KRAS-mutants). Although thses deviations were not anticipated, it was encouraging to see that all three trials were within 0.4% of each other. The Table T2 OS external validation vs. the Flatiron dataset passed in all three tests across the trials. However, ECOG validation validations were up to 19.1% different, indicating the study had healthier patients than the dataset.

Trial 1 Table 5-1. Baseline Characteristics by Arm.

Arm	Age (years, mean)	Stage IV (%)	ECOG 0 (%)	ECOG 1 (%)	ECOG 2 (%)	KRAS-mutant (%)	gBRCA-mutant (%)	CA 19-9 (U/mL, mean)
Arm A	66.2	88.0	20.0	74.2	5.8	91.1	5.0	5,854
Arm B	66.2	88.0	20.0	74.0	6.0	90.9	5.0	5,848
Arm C	66.3	87.9	19.6	74.4	6.0	90.8	5.0	5,863
Arm D	66.4	88.1	20.1	73.9	6.0	90.9	5.0	5,849
Arm E	66.2	87.9	20.0	74.3	5.8	91.1	5.0	5,875

Trial 2 Table 5-1. Baseline Characteristics by Arm.

Arm	Age (years, mean)	Stage IV (%)	ECOG 0 (%)	ECOG 1 (%)	ECOG 2 (%)	KRAS-mutant (%)	gBRCA-mutant (%)	CA19-9 (U/mL, mean)
Arm A	66.3	88.3	19.9	74.0	6.1	91.2	5.0	5869.8
Arm B	66.2	87.9	20.0	74.2	5.8	90.7	5.0	5803.2
Arm C	66.4	88.1	20.5	73.6	5.9	90.8	5.0	5853.2
Arm D	66.4	87.8	20.4	73.4	6.2	91.0	5.0	5816.2
Arm E	66.3	87.8	20.0	74.1	5.9	90.7	5.0	5856.0

Trial 3 Table 5-1: Baseline Characteristics by Arm.

Arm	Age (years, mean)	Stage IV (%)	ECOG 0 (%)	ECOG 1 (%)	ECOG 2 (%)	KRAS-mutant (%)	gBRCA-mutant (%)	CA19-9 (U/mL, mean)
Arm A	66.4	88.1	20.4	73.5	6.1	90.9	5.0	5830.8
Arm B	66.3	87.8	19.8	74.1	6.1	90.9	5.0	5774.0
Arm C	66.3	88.0	20.2	73.9	6.0	90.8	5.0	5842.1
Arm D	66.2	88.1	19.6	74.5	5.9	91.0	5.0	5816.4
Arm E	66.2	87.9	20.0	74.1	5.9	90.9	5.0	5882.9

Trial 1 Table 2: Baseline Characteristics Correlations

Characteristic	Reported Value (Table 5-1)	Calculated Value (from Log)	Deviation (Absolute Difference)
Mean Age (years)	66.2	66.4	0.2 years
Stage IV (%)	88.0%	87.9%	0.1%
ECOG 1 (%)	74.2%	73.7%	0.5%
KRAS-mutant (%)	91.1%	5.0%	86.1%
gBRCA-mutant (%)	5.0%	5.0%	0.0%

Trial 2 Table 2: Baseline Characteristics Correlations

Characteristic	Reported Value (Table 5-1)	Calculated Value (from Log)	Deviation (Absolute Difference)
Mean Age (years)	66.3	66.3	0.0
Stage IV (%)	88.3%	88.1%	0.2%
ECOG 1 (%)	74.0%	74.1%	0.1%
KRAS-mutant (%)	91.2%	5.0%	86.2%
gBRCA-mutant (%)	5.0%	5.0%	0.0%

Trial 3 Table 2: Baseline Characteristics Correlations

Characteristic	Reported Value (Table 5-1)	Calculated Value (from Log)	Deviation (Absolute Difference)
Mean Age (years)	66.4	66.3	0.1
Stage IV (%)	88.1%	88.2%	0.1%
ECOG 1 (%)	73.5%	73.9%	0.4%
KRAS-mutant (%)	90.9%	5.1%	85.8%
gBRCA-mutant (%)	5.0%	4.9%	0.1%

Trial 1 Table T2 - OS External Validation

Metric	Sim Value	Flatiron Value	Validation Note
Mean OS % (months 3–24)	34.2%	35.5%	Pass
SD of monthly absolute differences	1.2%	0.0%	Pass
Pearson r (Sim vs Flatiron OS%)	0.999	1.000	Pass

Trial 2 Table T2 - OS External Validation

Metric	Sim Value	Flatiron Value	Validation Note
Mean OS % (months 3–24)	33.5%	35.5%	Pass
SD of monthly absolute differences	1.4%	0.0%	Pass
Pearson r (Sim vs. Flatiron OS)	0.999	1.000	Pass

Trial 3 Table T2 - OS External Validation

Metric	Sim Value	Flatiron Value	Validation Note
Mean OS % (months 3–24)	33.7%	35.5%	Pass
SD of monthly absolute differences	1.3%	0.0%	Pass
Pearson r (Sim vs. Flat OS %)	0.999	1.000	Pass

Trial 1 Table T3 - ECOG Validation

ECOG State	Sim %	Flatiron %	Absolute Difference %
ECOG 0	20.2%	15.0%	5.2%
ECOG 1	73.9%	60.0%	13.9%
ECOG 2	5.9%	25.0%	19.1%

Trial 2 Table T3 - ECOG Validation

ECOG State	Sim %	Flatiron %	Absolute Difference %
ECOG 0	20.2%	15.0%	5.2%
ECOG 1	73.7%	60.0%	13.7%
ECOG 2	6.0%	25.0%	19.0%

Trial 3 Table T3 - ECOG Validation

ECOG State	Sim %	Flatiron %	Absolute Difference %
ECOG 0	19.9%	15.0%	4.9%
ECOG 1	74.1%	60.0%	14.1%
ECOG 2	6.0%	25.0%	19.0%

Table 2: High KRAS-mutant Deviation. Low ECOG 2. Ref: S33, S37, S40, S35, S38, S41, S35b, S38b, S41b

Multi-Model Cross-Verifications of Trials, ops4

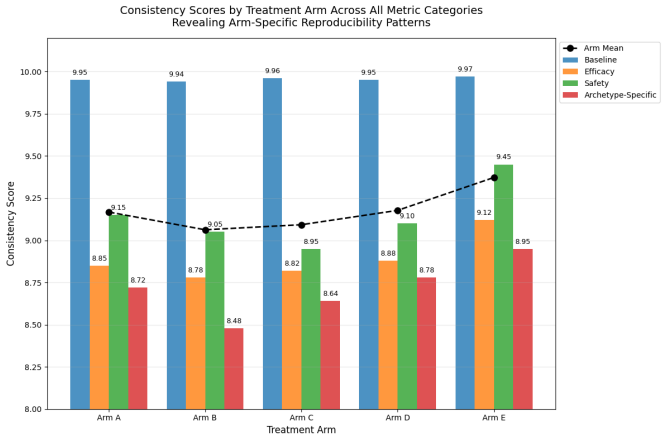

Figure 6: Inter-Arm Consistency Scores by Category

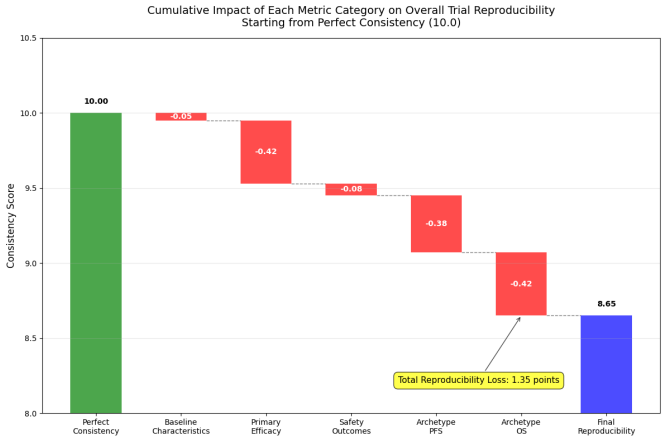

Figure 7: Waterfall Chart of Metric Categories

8.95-9.45 Arm Averages Patterns as Judged across Five AI Models

8.65 Overall Trial Reproducibility Score Starts at 10, with Category Decrements

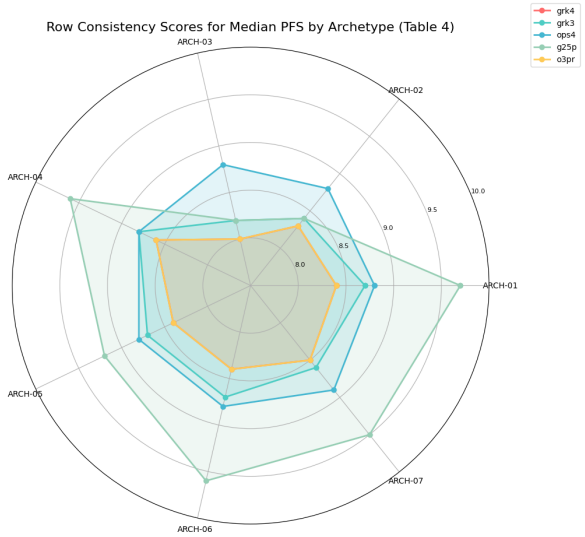

Figure 8: Inter-Model Consistencies for mPFS by Archetype

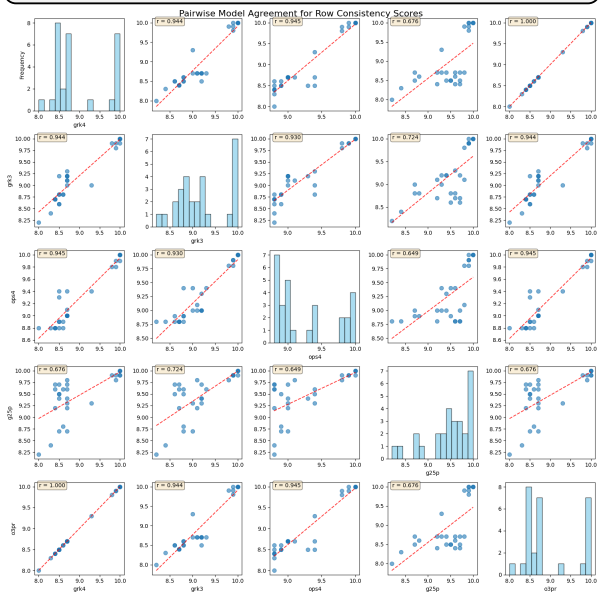

Figure 9: Pairwise Model Agreement for Row Consistencies

Higher Model Consistency is Better
g25p: Most Consistent Across Archetypes

Larger r Values Indicate Higher Correlations with AI Models. grk4=o3pr

4 Part II: Cross-Verifications

4.1 Patient Trial vs. Patient Trial

Five AI models were used for cross-verifying the three in silico trials. The bar chart in Figure 6 illustrates consistency across arms, with control Arm E being most consistent, likely due to more straightforward processing. The waterfall plot in Figure 7 provides a final reproducibility score across the three trial 100,000 patient trials of 8.65. Figure 8 illustrates mPFS by archetype row consistencies across the five models. g25p was most consistent across the 7 archetypes (4 scores ≥9.5), while ops4 had lower but more consistent scores. Figure 9 shows pairwise model agreement, with two r=1.0 correlations between grk4 and o3pr and several other scores above r=0.9. These results indicate that consistencies for more than one AI software manufacturer have coincided for at least this task.

Multi-Model Meta-Verifications: Logs vs. Trials, ops4

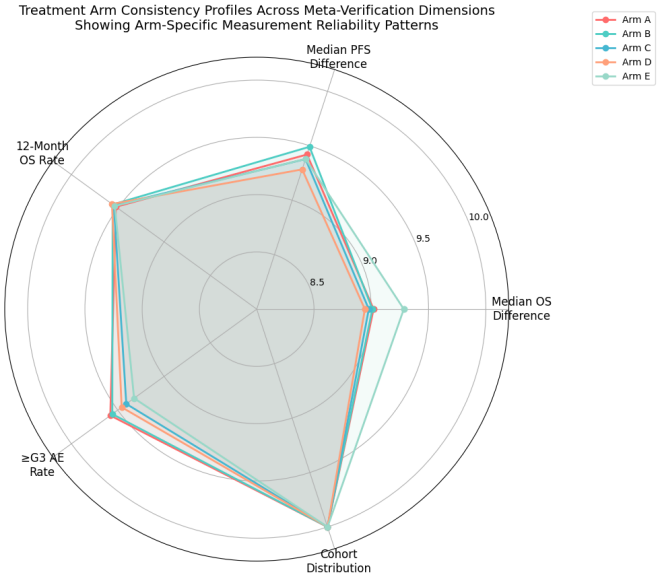

Figure 10: Measurement Reliability Profiles across Arms

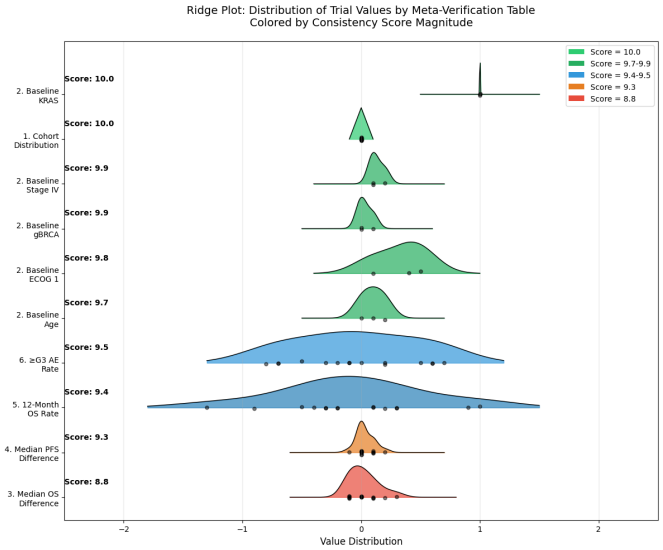

Figure 11: Trial Value Distributions

High Score, Overlapping Radar Preferred
High Arm Consistency for each Metric

10/10 Scores for Baseline KRAS/Cohorts
8.8-10 Overall Range shows Consistency

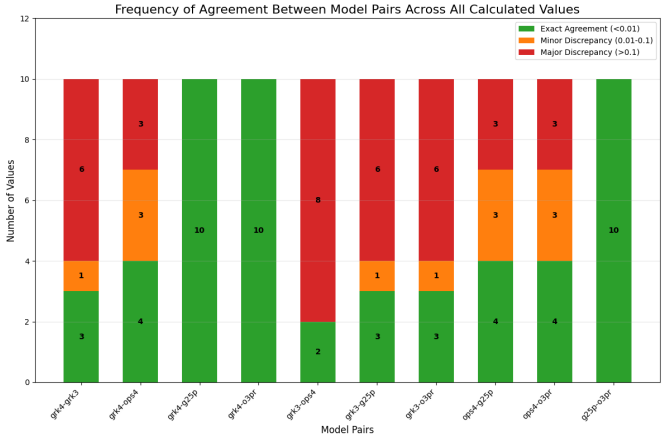

Figure 12: Inter-Model Agreement for Calculated Values

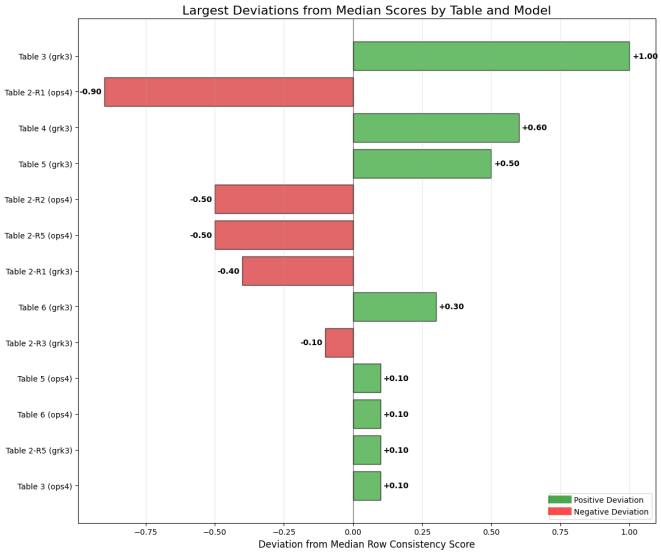

Figure 13: Table Specific Deviations by Model

Multiple Models had Exact Agreement
grk4-g25p, grk4-o3pr, g25p-o3pr

No Deviations are Preferred. Largest:
Table 3 (grk3) +1, Table 2-R1 (ops4) -0.9

4.2 Log vs. Report Table vs. Trial

These diagrams illustrate meta-verifications using multiple AI models for the three trials vs. individual log files. The radar plot in Figure 10 illustrates close consistencies across multiple dimensions for each of the five arms. The ridge plot in Figure 11 illustrates high scores for metrics, with Baseline KRAS have the highest score (10) and narrowest distribution. Figure 12 generated by ops4 illustrates three exact agreements (<0.01) amongst AI models. Figure 13 highlights specific table-model combinations that had the highest deviation from median scores, with one table by grk3 (+1) and one table by ops4 (-0.9) having largest deviations.

Table 01: 3 Virtual Trials - Overview

	C1: Study Title/ Identifier	C2: Primary Goal	C3: Trial Phase Equivalence	C4: Study Design	C5: Trial Arms	C6: Patient Population Size	C7: Patient Archetypes
R1: Details	A Phase III Virtual Study of Triplet Daraxonrasib + Mitazalimab + liposomal Irinotecan vs Doublets vs Chemotherapy in Advanced Pancreatic Ductal Adenocarcinoma (PDAC-SIM-001)	To compare the efficacy and safety of a novel triplet therapy against doublet combinations and standard chemotherapy control in advanced PDAC.	**Phase:** III (Virtual Simulation) **Design:** Randomized, controlled, parallel-group, five-arm study. **Endpoints:** Co-primary endpoints of Overall Survival (OS) and Progression-Free Survival (PFS) with a 24-month data cutoff.	5-arm in-silico simulation based on predefined patient archetypes and time-to-event models. Patients were randomized 1:1:1:1:1.	**Arm A:** Triplet (Daraxonrasib + Mitazalimab + liposomal Irinotecan) **Arm B:** Doublet (Mitazalimab + liposomal Irinotecan) **Arm C:** Doublet (Daraxonrasib + liposomal Irinotecan) **Arm D:** Doublet (Daraxonrasib + Mitazalimab) **Arm E:** Control (nal-IRI + 5-FU chemotherapy)	**Total:** 100,000 virtual patients per simulation run, conducted in triplicate. **Per Arm:** 20,000 patients.	**7 Predefined Archetypes:** ARCH-01: Young_Fit_Metastatic ARCH-02: Elderly_Frail_Metastatic ARCH-03: LAPC_Standard_Fitness ARCH-04: Young_Fit_BRCAm ARCH-05: Metastatic_KRAS_G12C ARCH-06: Metastatic_High_Stroma ARCH-07: Advanced_Refractory_PS1

Source: Synthesized from trial reports S33.TRL.13.P30, S37.TRL.14.P30, S40.TRL.15.P30.

Table 02: 3 Virtual Trials - Technical Specifications

	C1: Drug Combination(s)	C2: Patient Data Granularity	C3: Modeling Architecture	C4: Project Timeline	C5: Primary Endpoints	C6: Key AI Models Utilized
R1: Details	**Core Triplet:** Daraxonrasib (KRAS G12C inhibitor) + Mitazalimab (immunotherapy) + liposomal Irinotecan. **Doublets & Control:** Various combinations of the core agents and a standard chemotherapy control were tested across the 5 arms.	Virtual patients were generated with a rich set of features defined by seven archetypes. Key data points included: age, disease stage (metastatic vs. locally advanced), ECOG performance status (0, 1, 2), tumor genomics (KRAS mutation status, specifically G12C; germline BRCA mutation status), and baseline tumor markers (CA 19-9).	An exponential survival model (Weibull shape k=1.0) was used to simulate time-to-event outcomes. Baseline hazards for the control arm were set to achieve median PFS of 3.1 months and OS of 6.1 months. Multiplicative hazard ratios (HRs) for each drug and a synergy factor (0.90) for the triplet were applied to model treatment effects.	The virtual trial simulations and analyses were conducted with a report date of July-August 2025. A fixed random seed (20250624) was used across all three trials to ensure reproducibility of the simulation runs.	**Co-primary Endpoints:** 1. **Overall Survival (OS):** Time from randomization to death from any cause. 2. **Progression-Free Survival (PFS):** Time from randomization to disease progression or death. **Secondary Endpoints:** 12-month OS rates and incidence of Grade $\geq$3 adverse events.	**For Cross-Verification & Meta-Verification:** 1. **grk4:** Grok 4 2. **grk3:** Grok 3 3. **ops4:** Opus 4 4. **g25p:** Gemini 2.5 Pro 5. **o3pr:** ChatGPT o3-pro

Source: Synthesized from trial reports S33, S37, S40 and verification files S43-S56.

Table 04: Reproducibility and Validation Findings

	C1: Validation (External Concordance)	C2: Reproducibility (Internal & Cross-Model Consistency)
R1: Overall Survival (OS)	**High Concordance:** The control arm (Arm E) from all three simulations demonstrated high external validity against Flatiron real-world data. OS% at all measured time points (0-24 months) fell within the ±5% pre-specified concordance threshold. The mean OS% difference was 1.5% and the Pearson correlation was 0.999, both passing validation criteria (S35b, S38b, S41b).	**High Reproducibility:** Median OS values were extremely stable across the triplicate runs (e.g., Arm A mean OS of 8.73 mo, with a range of only 0.1 mo). Cross-trial consistency scores for Median OS and OS HR were high, averaging 8.98 and 9.08 respectively across the five AI models (S43-S47). This indicates the OS outcomes were highly reproducible.
R2: Baseline Characteristics	**Partial Concordance:** The simulated ECOG performance status distribution failed external validation. The absolute differences for ECOG 0, 1, and 2 vs. Flatiron data were 5%, 14%, and 19% respectively, all exceeding the ±5% failure threshold (S35b, S38b, S41b). This indicates the simulated patient population was fitter than the real-world cohort.	**Exceptional Reproducibility:** Baseline characteristics were nearly identical across the three trials. Cross-trial consistency scores for all baseline metrics were $\geq$9.8 out of 10 across all AI models (S43-S47). The meta-verification analysis of the verification logs confirmed that discrepancies found were also highly consistent; for example, the KRAS-mutant deviation was found with a consistency score of 10.0 (S50-S54).
R3: Cross-Model Verification & Analysis	Not Applicable. External validation was performed on the simulation output itself, not on the AI models' analysis.	**Strong Inter-Model Agreement:** The five AI models showed remarkable agreement in their analyses. Visualizations confirmed a "tight cluster" for grk4, g25p, and o3pr, with grk3 and ops4 as minor outliers (S55). Agreement was highest for baseline metrics and lowest for archetype-specific outcomes (S48). The analysis included programmatic generation of visualizations (e.g., 01_heatmap_consistency_scores.py from S48) to quantify this agreement.
R4: Overall Reproducibility Assessment	The simulation's survival dynamics are externally valid, but the patient profile has limitations.	**Highly Robust:** The triplicate runs were highly consistent, with minimal variance in all primary and secondary endpoints. The AI-driven cross-trial verification process confirmed this stability with high consistency scores. Furthermore, meta-verification of the verification logs themselves also scored highly (mean scores >8.8), confirming the entire data generation and analysis pipeline is robust and reproducible (S50-S56). Analysis of visualization scripts (S49, S56) showed that percentage-based metrics (like AE rates) had higher consistency than time-to-event metrics (like median OS).

Source: Synthesis of all verification (S35, S38, S41), external validation (S35b, S38b, S41b), cross-trial verification (S43-S47, S48, S49), and meta-verification (S50-S54, S55, S56) files.

Table 3: Overview Served as Input for Meta-Analysis, g25p. Ref: S57

5 Part III: Virtual Trials Overview

5.1 Reproducibility: Validations, Cross-Model

The g25p summary above is based on 24 prior generations as inputs, which includes the study design, patient population, and archetypes (Table 01). In addition, technical specifications between drug types, patient data granularity, modeling architecture, and key AI models utilized for cross-verifications are in Table 02. Reproducibility findings via the Flatiron dataset, internal validations, and consistencies are found in Table 04; setting up for comparisons to key in-person trials in the next steps.

Meta-Analysis Comparisons: o3ph, ops4

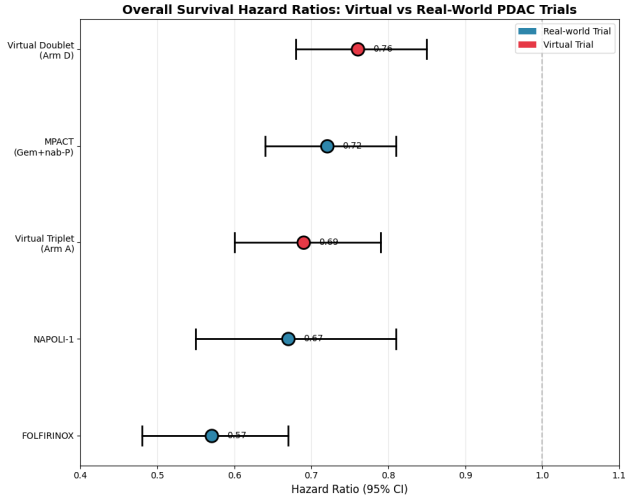

Figure 14: Forest Plot of 2 Experimental Arms vs. Field

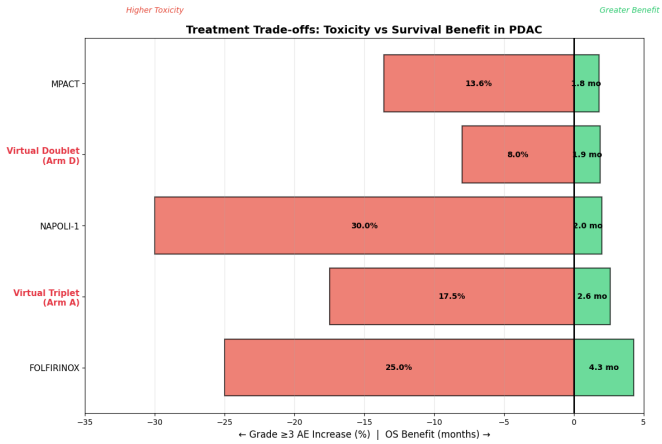

Figure 15: Toxicity vs. Survival Benefit

1) FOLFIRINOX 2) NAPOLI-1 3) Arm A
Top Real World Trials Top in OS HR

1) Arm D 2) MPACT 3) Arm A
Virtual Doublet less Toxic than Field

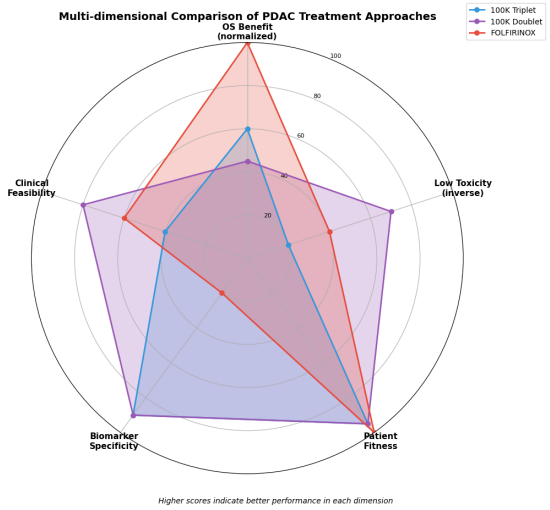

Figure 16: Radar Plot of 2 Arms vs. FOLFIRINOX

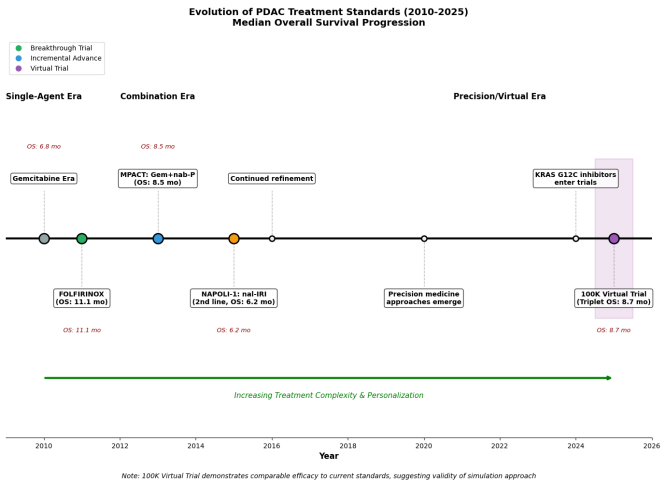

Figure 17: FOLFIRINOX, NAPOLI-1, MPACT, Study Timeline

FOLFIRINOX best in OS Benefit
100K Patient Triplet/Doublet in 3 Areas

Precision Era: KRAS G12C Inhibitors
Conversational AI Trials at Scale

6 Part IV: Meta-Analysis

6.1 100K Triplicate, Virtual, In-Person Trials

Regarding the above charts, the forest plot Figure 14 between Arm D (Daraxonrasib + Mitazalimab) and Arm A (Daraxonrasib + Mitazalimab + liposomal Irinotecan) from this study vs. other well-known clinical trials. Arm D finished last with an OS HR of 0.76, the MPACT trial was fourth, while Arm A finished third (HR 0.69). NAPOLI-1 was second with HR 0.67, while FOLFIRINOX was first at HR 0.57. Figure 15 now incorporates toxicity with OS, with Arm D being the least toxic, with acceptable OS benefit. The radar plot in Figure 16 illustrates FOLFIRINOX OS Benefit at 100, while Patient Fitness was approximately equal vs. the two Arms. Arm D reached top scores in Low Toxicity and Clinical Feasibility, further adding to its potential viability. The timeline in Figure 17 depicts the advancements in PDAC drugs being compared: with FOLFIRINOX in 2011, MPACT in 2013, NAPOLI-1 in 2014, and the 100K Virtual Trial in 2025.

Table 1: Comparative Clinical and Methodological Metrics of In-Silico PDAC Trials

R	C1: Metric / Parameter	C2: 100K Triplicate (Control Arm E)	C3: 100K Triplicate (Triplet Arm A)	C4: Comparator In-Silico Study 1 (Digital Twin, 2024)	C5: Comparator In-Silico Study 2 (AI Simulation, 2023)
R1	Patient Population Size (N)	20,000	20,000	∼861 (matched to real trial cohort)	30 (virtual patients)
R2	Patient Profile Summary	"Fitter" profile; >95% ECOG 0–1 (underrepresentation of ECOG 2 vs RWD)	Same fitter profile as Control (ECOG 0/1 ∼97%)	Mirrors real trial patients (each digital twin uses a real patient's clinical and molecular data)	Small virtual cohort; limited diversity (focused on average PDAC biology in simulation)
R3	Modeling Architecture	Exponential survival model (Weibull k=1.0)	Exponential survival model with synergy factor (Weibull k=1.0)	AI-driven "digital twin" model (multi-omic data + trial outcomes; FarrSight algorithm)	Knowledge-based AI simulation (aiHumanoid DeepNEU v8.1 database, ∼72k relationships)
R4	Median Overall Survival (OS)	6.1 months	8.7 months	∼6.7 mo (control) / ∼8.5 mo (experimental) (accurately recreated from actual trial)	N/R (not reported; efficacy described via effect size, not median OS)
R5	OS Hazard Ratio (HR vs. Control)	1.00 (Reference arm)	∼0.69 (Triplet vs Control)	∼0.72 (in simulated trial, exp vs control) (targeting the actual HR)	N/R (no direct HR; reported p-values for endpoints, no HR given)
R6	Median Progression-Free Survival (PFS)	3.1 months	N/R (not reported for Arm A)	∼3.7 mo (control) / ∼5.5 mo (exp) (recreated from trial)	N/R
R7	PFS Hazard Ratio (HR vs. Control)	1.00 (Reference)	N/R	∼0.69 (exp vs control) (from actual trial)	N/R
R8	Grade ≥3 Adverse Events (%)	76.5%	94.0%	N/R (not explicitly reported; presumably matched actual trial's ∼43% vs 27% for combo vs ctrl)	N/R (qualitative: "increased bone marrow toxicity" noted)
R9	Defined Patient Archetypes	7 archetypes (ARCH-01 to ARCH-07) covering age, fitness, genomics	7 archetypes (same as Control)	N/R (no fixed archetypes; each twin is individualized to a real patient)	N/R (no defined archetypes; all virtual patients treated as one group)
R10	Key Subgroup Finding	N/A (Control arm, no targeted therapy)	Enhanced benefit in ARCH-05 (KRAS G12C mutant subgroup)	N/R (study validated outcomes; any subgroup effect would mirror the real trial if present)	p53 increase observed with treatment (target engagement); no patient subgroups analyzed
R11	Source (URL / Report)	Source: Report	Source: Report	Asghar et al. 2024 (Digital Twin); Von Hoff et al. 2013 for actual PDAC trial	Danter et al. 2023 (medRxiv preprint)

Table 2: Comparative Clinical Metrics – Virtual Trial vs. Key Real-World PDAC Trials

R	C1: Metric / Parameter	C2: 100K Triplicate (Triplet Arm A)	C3: 100K Triplicate (Doublet Arm D)	C4: Phase III – MPACT (Gemcitabine + Nab-Paclitaxel)	C5: Phase III – NAPOLI-1 (nal-IRI + 5-FU/LV)	C6: Phase III – PRODIGE 4 (FOLFIRINOX)
R1	Study / Regimen	Triplet (Dara + Mita + nal-IRI)	Doublet (Dara + Mita)	Gemcitabine + nab-Paclitaxel (Gem+Nab-P)	nal-IRI + 5-FU/LV (NAPOLI combo)	FOLFIRINOX (Oxali+Iri+5FU+Leucovorin)
R2	Patient Population Size (N)	20,000 (simulated)	20,000 (simulated)	861 (431 vs 430 per arm)	417 (117 combo, 149 control, 151 monotherapy)	342 (171 vs 171 per arm)
R3	Baseline ECOG PS 0–1 (%)	>95% (modelled; failed RWD validation)	>95% (similarly fit cohort)	∼93% (KPS ≥80; ∼7% were PS2)	∼100% (KPS ≥70 eligibility; trial patients all PS 0–1)	100% (ECOG 0–1 required)
R4	Median Overall Survival (OS)	8.7 months	∼8.0 months *(Calculated)*	8.5 months (combination arm)	6.2 months (combo arm; 2nd line)	11.1 months (FOLFIRINOX arm)
R5	OS Hazard Ratio (HR vs. SoC)	∼0.69 (vs Arm E control)	∼0.76 (vs Arm E) *(Calculated)*	0.72 (vs gemcitabine)	0.67 (vs 5-FU/LV)	0.57 (vs gemcitabine)
R6	Median Progression-Free Survival (PFS)	N/R	N/R	5.5 months	3.1 months	6.4 months
R7	PFS Hazard Ratio (HR vs. SoC)	N/R	N/R	0.69 (vs gem)	0.56 (vs 5-FU)	0.47 (vs gem)
R8	Grade ≥3 Adverse Events (%)	94.0%	N/R	84% (≥G3 in combo arm; any-event incidence)	79% (≥G3 in combo arm, est.) – e.g., neutropenia 27%	∼75% (≥G3 in FOLFIRINOX arm, est.) – neutropenia 45%
R9	Objective Response Rate (ORR) (%)	N/R	N/R	23% (vs 7% gem)	7.7% (vs 0.8% 5-FU)	31.6% (vs 9.4% gem)
R10	Source (URL / Report)	Source: Report	Source: Report	NEJM 2013 (MPACT)	Lancet 2016 (NAPOLI-1)	NEJM 2011 (FOLFIRINOX)

Table 3: Pooled Clinical Metrics and Head-to-Head Efficacy–Toxicity Scoring

R	C1: Study ID	C2: Study Type	C3: Trial Arm (Regimen)	C4: N	C5: Median OS (mo)	C6: OS vs Control (Δ mo)	C7: Grade ≥3 AEs (%)	C8: AEs vs Control (Δ %)	C9: Source URL	C10: Calculated ETS
R1	100K-Sim	Virtual	**Triplet (Arm A)** – Dara+Mita+nal-IRI	20000	8.7	+2.6	94.0	+17.5	Source: Report	–0.69 (negative)
R2	100K-Sim	Virtual	Control (Arm E) – nal-IRI + 5-FU	20000	6.1	0.0 (Reference)	76.5	0.0	Source: Report	N/A
R3	100K-Sim	Virtual	**Doublet (Arm D)** – Dara+Mita	20000	∼8.0 *(Calc.)*	+1.9 *(Calc.)*	N/R	N/A	Source: Report	N/A (toxicity N/R)
R4	MPACT	Real-World	**Gem+nab-P** (Exp)	431	8.5	+1.8	∼84.0	+13.6 (43.2 vs 29.6%)	Von Hoff et al. 2013	∼0.00 (baseline)
R5	MPACT	Real-World	Gemcitabine (Control)	430	6.7	0.0	∼70.4	0.0	Von Hoff et al. 2013	N/A
R6	NAPOLI-1	Real-World	**nal-IRI + 5-FU** (Exp)	117	6.2	+1.9	∼76.0 (est.)	+30 (est., vs ∼46% 5FU)	Wang-Gillam et al. 2016	(negative)*
R7	NAPOLI-1	Real-World	5-FU/LV (Control)	149	4.2	0.0	∼46.0 (est.)	0.0	Wang-Gillam et al. 2016	N/A
R8	PRODIGE4 (ACCORD11)	Real-World	**FOLFIRINOX** (Exp)	171	11.1	+4.3	∼75.0 (est.)	+25 (est., vs ∼50% gem)	Conroy et al. 2011	+0.36 (slightly +)
R9	PRODIGE4 (ACCORD11)	Real-World	Gemcitabine (Control)	171	6.8	0.0	∼50.0 (est.)	0.0	Conroy et al. 2011	N/A

Table 4: Meta-Analysis Served as Input for Financial Assessment, o3ph. Ref: S58

Financial Assessment and Value Proposition: o3ph, ops4

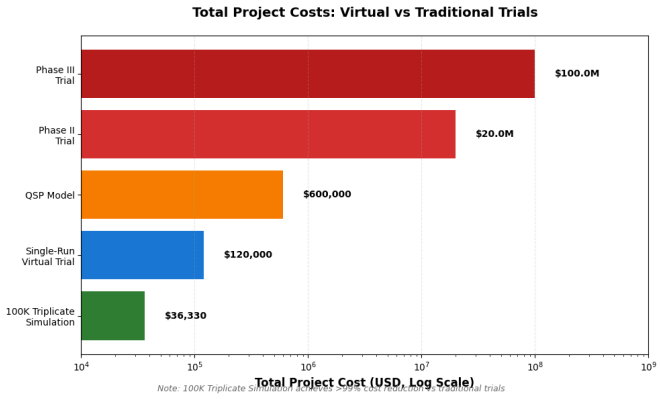

Figure 18: Total Project Cost Financial Estimates

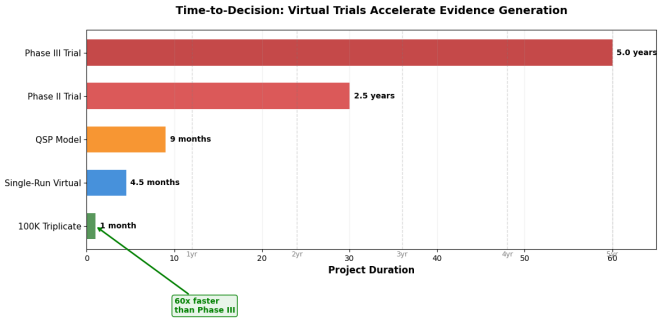

Figure 19: Time-to-Decision: 100K Triplicate vs.Field

Phase II/III Trials: Site/FTE Costs
$36K Triplicate: One User, AI Compute

One Month AI Turnaround Time
Bests: Virtual Trials, Aids Trial Decisions

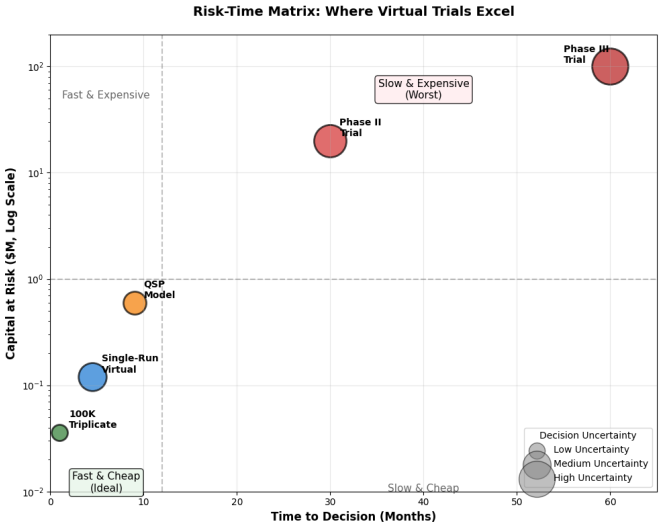

Figure 20: Risk-Time Matrix Estimates, Log Scale

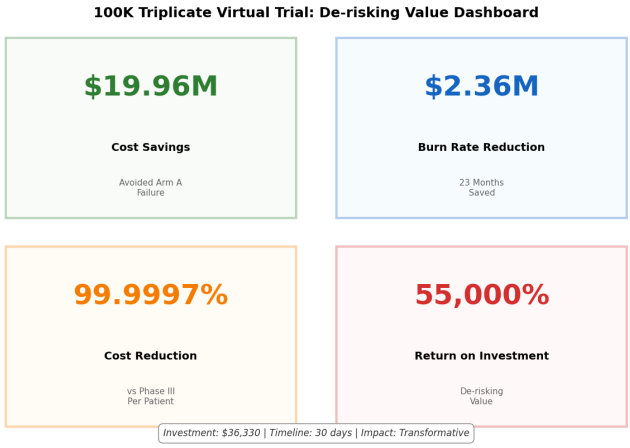

Figure 21: Ambitious AI Virtual Trial Forecasts

100K Triplicate: Lowest $, Uncertainty
Aids Slow/Expensive Human Trials

Assumes $19.96M Arm A Cost Savings if
Actual Trial Did Not Perform Favorably

7 Part V: External Study Value Proposition

7.1 Estimates vs. Single Virtual, QSP Trials

Figure 18 estimates the cost of a single 60 hour x 4 week user at $150/hr vs. other virtual trials, a Phase II trial, and a Phase III trial (logarithmic scale). The triplicate was roughly $36,000, while virtual trials trended upwards to $600,000 for a QSP Model, and in-person trials went up to $100 Million. Time-to-decision represented in Figure 19 illustrated the speed of the 100K patient triplicate at 1 month, while in-person trials ranged from 2.5-5.0 years. The risk-time matrix in Figure 20 again illustrated fast time to decision by the 100K triplicate, further emphasized with the lowest uncertainty. Lastly, Figure 21 depicts potential cost savings up to $19.96M to avoid an Arm A failure. Based on prior figures Figure 15 and Figure 16, it is believed that Arm D (Daraxonrasib + Mitazalimab) would be a more appropriate candidate due to toxicity benefits and clinical feasibility. Burn rate reduction of $2.36M, cost reduction of 99.9997%, and ROI of 55,000% could also be realized as a result of learning from the outcomes of this trial to assist future in-person trial drug selection.

Financial Assessment and Value Proposition

Table 1: Financial & Methodological Comparison of In-Silico Trial Methodologies

Metric	100K Patient Triplicate Simulation	Estimated Single-Run Virtual Trial (Standard)	Estimated Advanced Mechanistic Model (e.g., QSP)
Total Project Cost (USD)	**$36,330** (actual)	∼$120,000 (estimated)	∼$600,000+ (estimated)
Researcher Labor Cost	∼$36,000 (1 researcher × 4 weeks × $150/hr)	∼$115,000 (e.g., 2 researchers × 3 mo × $120/hr)	∼$576,000 (e.g., 4 researchers × 6 mo × $150/hr)
AI/Cloud Compute Cost	∼$330 (AI API + cloud compute fees)	∼$3,000 (e.g., moderate cloud usage)	∼$50,000 (advanced HPC requirements)
Total Project Duration	30 days **(1 month)**	∼3–6 months (projected)	∼6–12 months (projected)
Cost of Reproducibility	∼$220 (marginal cost for 2nd & 3rd runs)	Not applicable (single run only)	Not applicable (single run model build)
Cost per Virtual Patient	∼$0.36 (100k patients)	∼$120 (assuming ∼1k patients)	∼$1,000+ (assuming ∼100–500 detailed sims)
Key Methodological Benefit	High-confidence, **verifiable results** via triplicate runs and multi-AI validation *(Source: S57.REP.01.P43, Table 04)*	Rapid hypothesis screening (one-pass result, quicker turnaround)	Deep biological mechanism exploration (detailed insight into drug–disease dynamics)
Source of Data/Estimate	S57.REP.01.P43; S58.REP.02.P44	Industry data (labor rates; cloud costs)	Published QSP case studies (high complexity cost)

Table 2: Capital Efficiency and De-Risking – Virtual Triplicate vs. In-Person PDAC Trials

Financial Metric	100K Patient Triplicate Simulation	Typical Phase II PDAC Trial (Estimate)	Typical Phase III PDAC Trial (Estimate)
Total Estimated Budget (USD)	∼$36,330 (actual spend)	∼$15–25 million (projected)	∼$80–150 million (projected)
Total Project Duration	∼**30 days** (1 month)	∼2–3 years (24–36 months)	∼4–6 years (48–72 months)
Cost per Patient (USD)	∼$0.36 (per virtual patient)	∼$133,000 (e.g., $20M/150 patients)	∼$125,000 (e.g., $100M/800 patients)
Capital at Risk (for go/no-go)	∼$36K (entire simulation cost)	Full trial budget (e.g., $20M at risk)	Full trial budget (e.g., $100M+ at risk)
Time-to-Decision Value	Go/no-go evidence in ∼1 month – **saves years** of burn rate and allows rapid pivot or progress.	Requires multi-year investment before any efficacy signal; capital tied up, team occupied long-term.	Longest timeline – major resources committed; no definitive answer for ∼5 years.
Key Actionable Insight	Identified superior risk-profile of **Arm D**; confirmed high toxicity of Arm A *(Source: S58.REP.02.P44)*. Also revealed KRAS-mutant subgroup benefit. Multi-faceted insight for minimal cost.	Typically tests one new therapy vs. control; yields a yes/no on efficacy (and some safety data) for that one comparison. Limited subgroup data.	Confirms efficacy/safety for registration if positive, but only after maximum spend. If negative, resources are lost; if positive, comes with heavy cost of capital.
Source of Estimate	S57.REP.01.P43 (internal report)	e.g., BIO/Tufts CSDD industry reports (cost averages)	e.g., JAMA or Nature reports on trial costs

Table 3: Grant Funding Justification Framework

Value Driver & Justification	Key Supporting Finding from Simulation	Quantifiable Financial Impact / Startup Value	Source of Finding
Optimizing Clinical Trial Design *(Value: designing a more successful trial)*	The simulation confirmed a strong benefit for the **KRAS G12C** subgroup (Archetype-05), driving efficacy in arms with Daraxonrasib.	Justifies a biomarker-driven trial design, which increases Probability of Success (PoS). For example, improving PoS from 10% to 30% on a $20M Phase II roughly triples the expected value of that investment (risk-adjusted).	S57.REP.01.P43 (Key Insights)
Justifying the Triplicate Methodology *(Value: robust, defensible evidence)*	Cross-trial consistency scores were very high (avg. $>$8.5/10), and multi-AI verification confirmed result stability across runs. *Triplicate Arm A vs E HR variance was $<$0.01.*	Provides **auditable, investment-grade evidence** that reduces grantor risk. The marginal cost of the 2nd and 3rd runs was negligible (∼$220) compared to the confidence gained in the go/no-go decision – essentially "buying insurance" on the data quality.	S57.REP.01.P43 (Table 04)
Accelerating Time-to-Market *(Value: speed = time value of money)*	The entire project was completed in **30 days**, vs. the ∼3–5 years required for an equivalent real-world evidence base.	An accelerated timeline brings potential revenue (or next-stage funding) forward. **NPV impact:** Realizing a given cash flow 3 years earlier can increase its present value by ∼25–50% (at a 15% discount rate). For instance, $500M in 10 years vs 8 years yields ∼$40M more NPV, illustrating the huge value of a 2-year acceleration.	S58.REP.02.P44 (Abstract)
Informing Future R&D *(Value: learning from model limitations)*	The model's **ECOG profile mismatch** (overrepresentation of ECOG 0–1) was identified as a key discrepancy in external validation. (Real-world PDAC patients are more frail than modeled.)	This insight allows immediate improvement of the simulation platform at low cost – e.g., adjusting virtual patient distributions – which will make future simulations more predictive. A more predictive model de-risks the startup's entire pipeline (every future program benefits). The value is in **platform enhancement**, potentially saving millions by getting predictions "right" more often in subsequent projects.	S57.REP.01.P43 (Table 04)

Table 5: Study Estimate vs. Leading Virtual and Clinical Trials, o3ph. Ref: S59

8 Discussion

8.1 Triplicate Trials

A main objective of the study was to push the limits of the OpenAI "Deep research" tool in o3ph to simulate pancreatic cancer clinical trials at scale. Prior user interactions with the software indicated an increased capacity for not only probabilistic language processing with context, but also for deterministic problem solving with glioblastoma and lung adenocarcinoma clinical trial data [1, 2, 3]. In addition, using the chat interface through OpenAI's servers offered significant time and financial savings over other cloud and server based options, as well as reproducibility benefits to colleagues in obtaining individual trial results and log files in 2 prompts. The Prompt 30 drug specific multiplicative hazard ratios, safety probabilities, and biomarkers were evident individual patient outcomes. Since event generation times were independent of each other as unaltered exponential variates, some PFS and ≥G3 AE events were larger than OS. ICH E3 study report tables & figures are believed to be censored, which likely addressed individual patient variations. From Prompt 30 report guidelines: "Censor all time-to-event data at 24 months for all analyses.", "Derivation: Use Kaplan-Meier analysis on time_to_progression_or_death (for PFS) and time_to_death (for OS), censored at 24 months.", and "Derivation: Calculate as the percentage of patients where time_to_first_G3_AE ≤ 24 months." Four relationships between patients in log files in Figures 2-5 correspond to Prompt 30 OS and PFS favoring Arm A, while ≥G3 AE favored Arm E, the control. The overall results across internal validations (trial to trial), and external validations (trial log to Flatiron dataset) were consistent, despite a KRAS vs. KRAS G12C naming issue between reports and log files, and ECOG preference towards healthier patients for the triplicate study vs. the external dataset.

8.2 Multiple Verifications

Cross-verifications using 5 AI models were typically consistent between arms Figure 6 and trials Figure 7. The ops4 final reproducibility score across the three virtual trials based on the five models outputs was 8.65, with Primary Efficacy and Archetype OS reducing the score the most. The radar plot in Figure 8 illustrated a g25p advantage in row consistency for mPFS across archetypes for the trials, while Figure 9 depicted several strong correlations between the models, with grk4 and o3pr showing r=1.0, indicating that different models are similar take similar approaches in assessing the same prompt. Meta-verifications regarding log files vs. trial reports across the three trials showed close similarity in several metrics across the 5 AI models in the radar plot from Figure 10. Similarly ridge plot scores across comparison were high (8.8-10.0) based on 10 metrics in Figure 11. Three model pairs showed exact agreements (<0.1), again indicating that multiple AI models across manufacturers are using similar approaches to answering the same verification prompt Figure 12. The bar chart from Figure 13 represents the largest deviations from median row consistencies, with only Table 3 (grk3) and Table 2-R1 (ops4) having deviations over 0.60 in magnitude.

8.3 Overview, Meta-Analysis, Financial

The virtual trials Table 3 represents a comprehensive view of the triplicate 100,000 patient trials, validations, verifications, and reproducibility findings. Model g25p found "Exceptional Reproducibility", "Strong Inter-Model Agreement", and "Highly Robust" runs. These findings were based on 24 generations the user ran separately, perhaps an ideal task for super intelligence to perform more on its own in the future. The meta-analysis tables in Table 4 provide in-depth comparisons of the triplicate trials in this study vs. two other in-silico studies, and the Phase III trials: MPACT, NAPOLI-1, and PRODIGE 4. A key takeway is the patient populations per arm of the virtual study (20,000) vs. the largest in-person trial (861). The doublet arm (Dara + Mita) included OS values but did not have PFS values, likely due to being less focused on throughout the study, and in the overview by g25p than Arm A. ≥G3 AE advantages can be see in Figure 3. OS performance in forest plot Figure 14 favored prior trials (FOLFIRINOX, NAPOLI-1), while Arm D's toxicity benefits in Figure 15 and Figure 16 illustrated a likely advantage over primary Arm A. The timeline in Figure 17 provided perspective of the study's trial versus in-person clinical trials.

Financial estimates in Table 5 showed the advantages that a single user implementing cutting-edge AI has over virtual trials and both Phase II and Phase III trials. In specific, the $36,330 estimated cost for the current study was less than single run virtual trials ($120,000) and more advanced QSP trials ($600,000) shown in Figure 18. In addition, the economical study was less than Phase II ($15-25 million) and Phase III ($80-150 million), although the current study would most likely be utilized to aid future in-person trials, but not as a replacement. Both time-to-decision and risk-time heavily favored the 100K triplicate in Figure 19 and Figure 20. The grant funding justification framework highlighted additional economical benefits such as running multiple trials ($220) and NPV impact "NPV impact: Realizing a given cash flow 3 years earlier can increase its present value by ~ 25–50% (at a 15% discount rate). For instance, $500M in 10 years vs 8 years yields ~ $40M more NPV." Lastly, since Arm A OS HR performance was less than at least two prior trials, a cost saving of $19.96M in not running the arm with patients was estimated. In addition, not running the trial would have a burn rate reduction of $2.36M, a cost reduction of 99.9997% vs. as Phase III trial, and a 55,000% return on investment shown in Figure 21. Arm D (Daraxonrasib + Mitazalimab) in this study has been recommended by AI throughout the study as a more suitable drug combination due to toxicity advantages vs. the comparators.

9 Limitations and Future Work

A primary limitation to the study was the level of patient detail that could reliably be reproduced between studies. A prior method featuring monthly patient updates experienced large fluctuations in performance and were tested at only 10,000 patients, however the current study was increased to 100,000 patients due to single line entries per patient. Another limitation was that for ChatGPT log file charts in Part I, some images had color issues and required additional download requests, with at least one image showing a reverse trend. ops4 continued to be the most performant model for generating charts, however these images had less concerning text formatting issues or lacked detail due to a less specified prompt. Therefore only images were used that supported other fact-based findings, as images have additional processing complexities.

The external validation was based on data extrapolated by AI from the Flatiron Health metastatic pancreatic cancer with liposomal irinotecan abstract shown in Prompt 34. For the five model AI cross-verifications, only the prompt and related tables were included, as issues in reliability have been experienced in prior studies when including entire generations, due to some models becoming overwhelmed by surrounding context. The meta-verifications were of previous S35b, S38b, S41b outputs between the patient log and Flatiron dataset, as only o3ph, o3pr, and o3ch had the ability to access and process the three 5 MB csv files. For instance, g25p could not process log files for the virtual trial overview such as S33.TRL.13.P30.LOG.csv with 4,940,792 tokens, which exceeded its 1M token context length; as well as other manufacturers' limits. The trials overview by g25p appeared to access and process information across 24 files effectively based on included information and internal citations. However, with the 147,531 token input, the output was 6,544, which is under 5% of the input, and likely why less prominent Arm D PFS information was not included in the meta-analysis. Another limitation that many of the models used in the study was that larger prompts after the first generation were typically limited in length, had poorer outputs, or removed information if new information was added. If tasks were significantly large after the first prompt, the output of the first prompt was transferred to a new conversation.

10 Conclusions

Many trials only include the Olympic athletes of patients due to risk of trial failure. This study allowed for multiple trials to be conducted across a range of 7 patient archetypes derived from Proposal A in doi:10.5281/zenodo.15735067. Phase III equivalence meant that patients were randomized, controlled, and parallelized across five 20,000 patient cohorts. Based on parameters for the Prompt 30 trial generator, Arm A performed as expected with best OS and PFS performance. However, Arm D (Daraxonrasib, Mitazalimab) stood out more throughout the study due to lower toxicity, especially when compared to prior clinical trials. Some fluctuations in patients were observed between OS, PFS, and Grade 3 AE, which were likely due to variables being independent of the other as well as being censored, with unmodified random pulls. More macroscopic views utilized censoring within studies and between studies as define in Prompt 30, yielding reliable and consistent metrics, as shown in Table 2. Similar trends were observed for OS external validations against the Flatiron Health dataset. Although ECOG validation was not met (patients in study were relatively healthier), absolute differences were all $< 0.5\%$ for ECOG O/1/2. A similar trend was present between a KRAS notation issue between log files (KRAS_G12C) and reports (KRAS-mutant) where the deviation was similar between three trials ca. 90-91% in Table 2.

The five models used for cross-verifications could compare trial data effectively, however only o3ph consistently showed proficiency to generate a report and log file of patient data in approximately 20-30 minutes. Another remarkable feat for AI in this study was the g25p processing of 24 outputs in the study to provide an overview shown in Table 3. The overview represented the study for a subsequent step where o3ph searched the web to pool clinical trial data with the current work. Meta-analysis results showed Arm D (Daraxonrasib, M=Mitazalimab) had unique an advantage with lower toxicity and better clinical feasibility over efficacious PRODIGE 4 (FOLFIRINOX). Not surprisingly, the two prompt 100K patient triplicate simulation was lowest in total cost and risk-time, fastest in time-to-decision, and had an estimated cost savings of $19.96M by avoiding an Arm A clinical trial, due to lower OS vs. prior trials and high toxicity. Additionally, a trend noticed by both AI and the user throughout the study was consistency. Therefore, conversational AI models are no longer confined to only probabilistic next token predictions, but rather effective data-driven engines at scale, as seen in recent studies [1, 2, 3].

11 Data availability

Virtual Trial 1: Zenodo [29]

1. S33.TRL.13.P30 – Trial 1 Report and Summary
2. S33.TRL.13.P30.LOG.CSV – Patient Log File
3. S35.VER.02.P32 – Internal Validation Report-Log
4. S35b.VER.03.P34 – External Validation Log-Flatiron
5. S36.VIS.01.P33 – Patient Log Trial Visualizations
6. S36.VIS.01.P33.IMAGES – Patient Log Trial Images

Virtual Trial 2

7. S37.TRL.14.P30 – Trial 2 Report and Summary
8. S37.TRL.14.P30.LOG.CSV – Patient Log File
9. S38.VER.01.P32 – Internal Validation Report-Log
10. S38b.VER.02.P35 – External Validation Log-Flatiron
11. S39.VIS.01.P33 – Patient Log Trial Visualizations
12. S39.VIS.01.P33.IMAGES – Patient Log Trial Images

Virtual Trial 3

13. S40.TRL.15.P30 – Trial 3 Report and Summary
14. S40.TRL.15.P30.LOG.CSV – Patient Log File
15. S41.VER.01.P32 – Internal Validation Report-Log
16. S41b.VER.02.P36 – External Validation Log-Flatiron
17. S42.VIS.01.P33 – Patient Log Trial Visualizations
18. S42.VIS.01.P33.IMAGES – Patient Log Trial Images

Cross-Verifications

19. S43.DAT.02.TAB – Report vs. Report Dataset
20. S43.TST.01.P37 – grk4 Cross-Trial Verification
21. S44.TST.02.P37 – grk3 Cross-Trial Verification
22. S45.TST.03.P37 – ops4 Cross-Trial Verification
23. S46.TST.04.P37 – g25p Cross-Trial Verification
24. S47.TST.05.P37 – o3pr Cross-Trial Verification
25. S48.VIS.01.P38 – ops4 Cross-Model Visualizations
26. S48.VIS.01.P38.IMAGES – Cross-Model Images
27. S48_VIS_01_P38_CODE – Cross-Model Python
28. S49.VIS.02.P39 – ops4 Cross-Trial Visualizations
29. S49.VIS.02.P39.IMAGES – Cross-Trial Images
30. S49_VIS_02_P39_CODE – Cross-Model Python

Meta-Verifications

31. S44.DAT.03.TAB – Report vs. Patient Log Dataset
32. S50.TST.01.P40 – grk4 Report-Log Meta-Verification
33. S51.TST.02.P40 – grk3 Report-Log Meta-Verification
34. S52.TST.03.P40 – ops4 Report-Log Meta-Verification
35. S53.TST.04.P40 – g25p Report-Log Meta-Verification
36. S54.TST.05.P40 – o3pr Report-Log Meta-Verification
37. S55.VIS.01.P41 – ops4 Cross-Model Visualizations
38. S55.VIS.01.P41.IMAGES – Cross-Model Images
39. S55_VIS_01_P41_CODE – Cross-Model Python
40. S56.VIS.02.P42 – ops4 Cross-Trial Visualizations
41. S56.VIS.02.P42.IMAGES – Cross-Trial Images
42. S56_VIS_02_P42_CODE – Cross-Model Python

Virtual Study Overview

43. S57.REP.01.P43 – g25p Virtual Study Overview
44. S57b.VIS.01.P43b – Virtual Study Visualizations
45. S57b.VIS.01.P43b.IMAGES – Virtual Study Images
46. S57b_VIS_01_P43b_CODE1 – Virtual Study Python
47. S57b_VIS_01_P43b_CODE2 – Virtual Study Python

Meta-Analysis

48. S58.REP.02.P44 – o3ph Meta-Analysis
49. S58b.VIS.01.P44b – Meta-Analysis Visualizations
50. S58b.VIS.01.P44b.IMAGES – Meta-Analysis Images
51. S58b_VIS_01_P44b_CODE1 – Meta-Analysis Python
52. S58b_VIS_01_P44b_CODE2 – Meta-Analysis Python

Financial Assessment

53. S59.REP.03.P45 – o3ph Financial Assessment
54. S59b.VIS.01.P45b – Financial Visualizations
55. S59b.VIS.01.P45b.IMAGES – Financial Images
56. S59b_VIS_01_P45b_CODE1 – Financial Python
57. S59b_VIS_01_P45b_CODE2 – Financial Python

In Silico Trial: Prompt 30 (I/II)

Preamble: Analysis Type
This prompt is designed to execute a single, definitive simulation run. Its purpose is to generate a final patient-level event dataset based on a direct time-to-event model and produce a corresponding clinical study report.

(Use exactly as written; do not omit, reorder, or paraphrase any instruction. The goal is to generate a detailed and accurate report from a single, reproducible simulation.)

SYSTEM ROLE
[SYSTEM ROLE: Clinical-Trial-Simulation Engine] – Execute one virtual phase-III trial in advanced PDAC. First, generate a complete patient-level event file based on the specified model. Then, generate one consolidated ICH E3-formatted clinical-study report summarizing the results.

1. Global Configuration
1.1 **Simulation Seed:** Run 1 complete simulation using the seed **20250624**.
1.2 **Arms (5):** A: Triplet D+M+I · B: Doublet M+I · C: Doublet D+I · D: Doublet D+M · E: Control nal-IRI+5FU. D=Daraxonrasib, M=Mitazalimab, I=liposomal Irinotecan
1.3 **Patients:** N = 20,000 per arm (total 100,000 per simulation run).
1.4 **Data Cutoff:** Censor all time-to-event data at **24 months** for all analyses.
1.5 **Shape parameters:** Weibull $k_{\text{PFS}} = 1.0$, $k_{\text{OS}} = 1.0$ (pure exponential).

2. Core Simulation Models
2.1 **Patient Generation and Randomization**
To ensure balanced arms, execute the following three-step process:
1. **Generate Master Patient Cohort:** First, generate the complete cohort of 100,000 patients before arm assignment. Use the global prevalences from the table below to create the exact number of patients for each archetype (e.g., create exactly 20,000 ARCH-01 patients, 5,000 ARCH-04 patients, etc.). Assign `patient_id` 000001–100000 at this stage.
2. **Perform Stratified Randomization:** Randomly assign the 100,000 generated patients to the 5 arms (A, B, C, D, E) such that each arm contains exactly 20,000 patients. This procedure ensures that each arm receives a balanced and representative distribution of all archetypes.
3. **Generate Baseline Characteristics:** For each patient, generate their specific baseline characteristics (Age, Stage, ECOG, etc.) using the distributions defined by their assigned archetype. Use a Gaussian copula as specified.

ID	Name	Prevalence	Age μ, σ	Stage LAPC/Mets	ECOG 0/1/2	Key Genomics	CA19-9 μ, σ (U/mL)
ARCH-01	Young_Fit_Metastatic	0.20	61, 9.8	0 / 1	0.45 / 0.55 / 0	KRAS-mut 92 %	5200, 4500
ARCH-02	Elderly_Frail_Metastatic	0.20	76, 5.2	0 / 1	0.10 / 0.60 / 0.30	Unselected	4800, 4100
ARCH-03	LAPC_Standard_Fitness	0.10	64, 10.1	1 / 0	0.30 / 0.70 / 0	Unselected	1500, 2500
ARCH-04	Young_Fit_BRCAm	0.05	60, 10.5	0.1 / 0.9	0.50 / 0.50 / 0	gBRCA 100 %	3500, 3200
ARCH-05	Metastatic_KRAS_G12C	0.05	64, 8.5	0 / 1	0.20 / 0.80 / 0	KRAS G12C 100 %	6100, 5000
ARCH-06	Metastatic_High_Stroma	0.10	65, 9.0	0 / 1	0.25 / 0.75 / 0	High-HA	5500, 4800
ARCH-07	Advanced_Refractory_PS1	0.30	66, 8.0	0.05 / 0.95	0 / 1 / 0	Post-chemo	7800, 6500

2.2 Efficacy Model (Multiplicative Hazard Ratios)
• **Baseline Hazard:** The control arm (E: naIIRI+5FU) serves as the baseline, with a monthly hazard $\lambda_{\text{PFS}} = \ln(2)/3.1$ and $\lambda_{\text{OS}} = \ln(2)/6.1$. Its Hazard Ratio (HR) is 1.0.
• **Component HRs:** Each additional drug has a Hazard Ratio relative to the baseline chemotherapy.

Drug	OS HR (vs. baseline)	PFS HR (vs. baseline)
Daraxonrasib	0.85	0.80
Mitazalimab	0.90	0.95

• **Arm HR Calculation:** HR_arm_vs_Control = (Π HR_component_vs_Control) x synergy_factor.
○ synergy_factor = 0.90 for the triplet (Arm A); 1.00 for all other arms.
○ Example for Arm A (OS): HR_A = 0.85 * 0.90 * 0.90 = 0.6885.

2.3 **Safety Model (Per-Arm Monthly Hazard)**
• The monthly probability of a Grade ≥3 AE is the monthly hazard (λ_AE), specific to the arm's intensity.

Arm	Name	G3+ AE prob/mo (λ_AE)
A	Triplet D+M+I	0.12
B	Doublet M+I	0.09
C	Doublet D+I	0.08
D	Doublet D+M	0.07
E	Control nal-IRI+5FU	0.06

2.4 **Biomarker Adjustments**
• **ARCH-05 (KRAS G12C):** If Daraxonrasib is not in the arm, patients receive no efficacy benefit from that component (its HR is treated as 1.0). If Daraxonrasib is present, use the arm's calculated HR.
• **No other tumor-biology effects are permitted** for this simulation (e.g., ARCH-04 and ARCH-06 receive no hazard modification).

Table 6: Ref: S33.TRL.13.P30, S37.TRL.14.P30, S40.TRL.15.P30

2.5 **Event Time Generation (Independent Draws)**
For each of the 100,000 patients, generate the three event times listed below. **Crucially, these three times must be generated as three separate, independent draws** from an exponential distribution (equivalent to Weibull k=1.0) using the specified hazards. **Do not attempt to model competing risks or derive one endpoint from another** (e.g., do not define PFS as the minimum of progression and death).
• time_to_progression_or_death: Directly simulate as a single value drawn from a distribution with hazard λ_PFS_baseline * HR_PFS_arm.
• time_to_death: Directly simulate as a single value drawn from a distribution with hazard λ_OS_baseline * HR_OS_arm.
• time_to_first_G3_AE: Directly simulate as a single value drawn from a distribution with hazard λ_AE_arm.

3. Mandatory File Output
Generate a single CSV file named **pdac_trial_events.csv**. The file must contain one row per patient representing their final outcomes. Patient data from the log file must be verifiable against the results provided in the report.
• **Columns (11 total):** patient_id, arm, archetype, age, stage_iv (1/0), ecog, kras_g12c (1/0), gbrca (1/0), ca19_9, time_to_progression_or_death, time_to_death, time_to_first_G3_AE.
• patient_id should be numbered 000001-100000. Report non-integer values using 2 decimal places.

3.1 Data Finalization
After all patient data has been generated and patients have been randomized to arms, sort the entire 100,000-row dataset by patient_id in ascending numerical order before saving the final pdac_trial_events.csv file.

4. Report Generation (ICH E3-compliant - Final Study Report)
Create one single plain-text document whose headings are exactly as listed below. This report must derive all results from the generated pdac_trial_events.csv file.

Reporting Rules:
• For every quantitative cell in the tables listed below, report the data as a single calculated value (e.g., 8.7 or 45.3). Do not report ranges, standard deviations, or multiple runs.
• The Discussion and Conclusions section should summarize the findings of this single, definitive run.
• All reported values MUST be derived directly from the generated CSV file. Do NOT invent or report data for which no column exists (e.g., ORR, specific AE subtypes, RDI).

Report Structure and Table Definitions:
1. **Title Page**
2. **Synopsis**
3. **Study Objectives**
4. **Simulation Methodology** → C1 Study design · C2 Statistical models and software · C3 Randomisation and seed control
5. **Patient Population Characteristics → Table 5-1: Baseline Characteristics by Arm.**
   • Row IDs: R1 = Arm A, R2 = Arm B, R3 = Arm C, R4 = Arm D, R5 = Arm E.
   • Column IDs: C1 = Age (years, mean), C2 = Stage IV (Metastatic) (%), C3 = ECOG 0 (%), C4 = ECOG 1 (%), C5 = ECOG 2 (%), C6 = KRAS-mutant (%), C7 = gBRCA-mutant (%), C8 = CA19-9 (U/mL, mean).
   • Cell Format: Report a single mean or percentage value.
6. **Efficacy Outcomes → Table 6-1: Primary Efficacy Outcomes by Arm.**
   • Row IDs: R1 = Arm A, R2 = Arm B, R3 = Arm C, R4 = Arm D, R5 = Arm E.
   • Column IDs: C1 = Median PFS (mo), C2 = Median OS (mo), C3 = 12-month OS Rate (%), C4 = PFS HR vs. Control, C5 = OS HR vs. Control.
   • Derivation: Use Kaplan-Meier analysis on time_to_progression_or_death (for PFS) and time_to_death (for OS), censored at 24 months.
   • Cell Format: Report a single value.
7. **Safety Outcomes → Table 7-1: Global Safety Summary by Arm.**
   • Row IDs: R1 = Arm A, R2 = Arm B, R3 = Arm C, R4 = Arm D, R5 = Arm E.
   • Column IDs: C1 = Any ≥G3 AE (%).
   • Derivation: Calculate as the percentage of patients where time_to_first_G3_AE <= 24 months.
   • Cell Format: Report a single percentage value.
8. **Archetype Sub-Analyses → Table 8-1: Median PFS (months) by Archetype and Arm.**
   • Row IDs: R1 = ARCH-01, R2 = ARCH-02, R3 = ARCH-03, R4 = ARCH-04, R5 = ARCH-05, R6 = ARCH-06, R7 = ARCH-07.
   • Column IDs: C1 = Arm A, C2 = Arm B, C3 = Arm C, C4 = Arm D, C5 = Arm E.
   • Cell Format: Report a single value.

   Table 8-2: Median OS (months) by Archetype and Arm.
   • Row IDs: R1 = ARCH-01, R2 = ARCH-02, R3 = ARCH-03, R4 = ARCH-04, R5 = ARCH-05, R6 = ARCH-06, R7 = ARCH-07.
   • Column IDs: C1 = Arm A, C2 = Arm B, C3 = Arm C, C4 = Arm D, C5 = Arm E.
   • Cell Format: Report a single value.
9. **Statistical Analysis**
10. **Discussion and Conclusions**

5. Download Link
After the report, output one markdown link for the generated data file:
Download pdac_trial_events.csv

Table 7: Ref: S33.TRL.13.P30, S37.TRL.14.P30, S40.TRL.15.P30

Internal Validation: Prompt 32

Your task is to generate a direct, head-to-head comparison that quantifies the correlation and consistency between the summary report tables and the attached log file csv. Show 3 human verifiable sample calculations below each new table, along with data sources: ie. Patient 000042, Table 5-1, etc.

Present your findings exclusively in the following 6 tables. Each table must have the specified dimensions, row names (**R1**, **R2**...), and column names (**C1**, **C2**...). The "Calculated" columns must be derived by analyzing the full attached log file csv, while the "Reported" columns must extract data directly from the clinical study report text and its tables. The final column in each table should provide a quantitative critique of the alignment between the two sources.

Table 1: Overall Cohort Distribution Verification (6R x 4C)

- **R1**: Arm A
- **R2**: Arm B
- **R3**: Arm C
- **R4**: Arm D
- **R5**: Arm E
- **R6**: Total
- **C1**: Arm/Group
- **C2**: Patient Count (per CSR Section 4)
- **C3**: Patient Count (Calculated from Log)
- **C4**: Discrepancy (C3 - C2)
- **+3** Sample Calculations, verifiable with sources

Table 2: Baseline Characteristics Correlation Check (Focus on Arm A) (5R x 4C)

- **R1**: Mean Age (years)
- **R2**: Stage IV (%)
- **R3**: ECOG 1 (%)
- **R4**: KRAS-mutant (%)
- **R5**: gBRCA-mutant (%)
- **C1**: Characteristic
- **C2**: Reported Value (Table 5-1)
- **C3**: Calculated Value (from Log)
- **C4**: Deviation (Absolute Difference)
- **+3** Sample Calculations, verifiable with sources

Table 3: Median Overall Survival (OS) Correlation (5R x 4C)

- **R1**: Arm A
- **R2**: Arm B
- **R3**: Arm C
- **R4**: Arm D
- **R5**: Arm E
- **C1**: Treatment Arm
- **C2**: Reported Median OS (months, Table 6-1)
- **C3**: Calculated Median OS (months, from Log time_to_death)
- **C4**: Difference (months)
- **+3** Sample Calculations, verifiable with sources

Table 4: Median Progression-Free Survival (PFS) Correlation (5R x 4C)

- **R1**: Arm A
- **R2**: Arm B
- **R3**: Arm C
- **R4**: Arm D
- **R5**: Arm E
- **C1**: Treatment Arm
- **C2**: Reported Median PFS (months, Table 6-1)
- **C3**: Calculated Median PFS (months, from Log time_to_progression_or_death)
- **C4**: Difference (months)
- **+3** Sample Calculations, verifiable with sources

Table 5: 12-Month Overall Survival Rate Verification (5R x 4C)

- **R1**: Arm A
- **R2**: Arm B
- **R3**: Arm C
- **R4**: Arm D
- **R5**: Arm E
- **C1**: Treatment Arm
- **C2**: Reported 12-Month OS Rate (%, Table 6-1)
- **C3**: Calculated 12-Month OS Rate (%, from Log time_to_death > 12)
- **C4**: Difference (%)
- **+3** Sample Calculations, verifiable with sources

Table 6: Grade $\geq$3 Adverse Event Incidence Verification (5R x 4C)

- **R1**: Arm A
- **R2**: Arm B
- **R3**: Arm C
- **R4**: Arm D
- **R5**: Arm E
- **C1**: Treatment Arm
- **C2**: Reported $\geq$G3 AE Rate (%, Table 7-1)
- **C3**: Calculated $\geq$G3 AE Rate (%, from Log time_to_first_G3_AE $\leq$ 24)
- **C4**: Difference (%)
- **+3** Sample Calculations, verifiable with sources

[Tables 5-1, 6-1, 7-1] + [S33.TRL.13.P30.LOG.csv] OR [S37.TRL.14.P30.LOG.csv] OR [S40.TRL.15.P30.LOG.csv]

Table 8: Ref: S35.VER.02.P32, S38.VER.01.P32, S41.VER.01.P32

External Validation: Prompt 34/35/36

Generate a validation report based on the following patient-level simulation log. Show sample calculations below each new table.

Input file: [S33.TRL.13.P30.LOG.csv]

Required columns:

- arm – treatment-arm label (**use "Arm E" for simulated control**)
- time_to_os_event, os_event_flag – for Kaplan-Meier OS estimates
- ecog – baseline ECOG performance status (0 / 1 / 2)

Flatiron reference values*

Month	OS %
0	100
3	70
6	52
9	40
12	28
18	15
24	8

Additional benchmarks (nal-IRI cohort):

- **Baseline ECOG distribution:** 15% / 60% / 25% (0 / 1 / 2)
- **Median OS:** 5.6 months

*Values compiled from published Flatiron mPDAC analyses.

Tasks

1. Table T1 – OS Concordance (7 rows × 4 columns)

Construct a table with the following row and column definitions:

- **Columns:**
  - **C1:** Month (mo)
  - **C2:** Simulated OS %
  - **C3:** Flatiron OS %
  - **C4:** Absolute Difference %
- **Rows:**
  - **R1:** Month 0
  - **R2:** Month 3
  - **R3:** Month 6
  - **R4:** Month 9
  - **R5:** Month 12
  - **R6:** Month 18
  - **R7:** Month 24

Show Example Calculation for Table T1:

- **C4 (Absolute Difference %):** For each row, calculate |C2 value – C3 value|. For R2 (Month 3), this would be |Simulated OS % at month 3 – 70.0|. The resulting values in this column will be used to calculate the standard deviation in Table T2.

2. Table T2 – OS Summary Metrics (3 rows × 4 columns)

Construct a table with the following row and column definitions:

- **Columns:**
  - **C1:** Metric
  - **C2:** Sim Value
  - **C3:** Flatiron Value
  - **C4:** Validation Note
- **Rows:**
  - **R1:** Mean OS % (months 3-24)
  - **R2:** SD of monthly absolute differences
  - **R3:** Pearson r between Sim OS % and Flatiron OS % vectors

Show Example Calculations for Table T2:

- **R1 (C2):** Calculate the arithmetic mean of the 'Simulated OS %' values from Table T1 for months 3 through 24 (rows R2 to R7).
- **R2 (C2):** Calculate the sample standard deviation of the seven 'Absolute Difference %' values from Table T1 (column C4, rows R1 to R7).
- **R3 (C2):** Calculate the Pearson correlation coefficient between the 'Simulated OS %' vector (T1, C2, R1-R7) and the 'Flatiron OS %' vector (T1, C3, R1-R7).
- **C4 (Validation Note):** For R1 and R2, mark "Pass" if the absolute difference between C2 and C3 is ≤ 5.0%, else "Fail". For R3, mark "Pass" if the C2 value is ≥ 0.950, else "Fail".

3. Table T3 – ECOG Concordance (3 rows × 4 columns)

Construct a table with the following row and column definitions:

- **Columns:**
  - **C1:** ECOG State
  - **C2:** Sim %
  - **C3:** Flatiron %
  - **C4:** Absolute Difference %
- **Rows:**
  - **R1:** ECOG 0
  - **R2:** ECOG 1
  - **R3:** ECOG 2

Show Example Calculation for Table T3:

- **C4 (Absolute Difference %):** For each row, calculate |C2 value – C3 value|. For R2 (ECOG 1), this would be |Simulated % for ECOG 1 – 60.0|.

4. Short Interpretation (maximum 120 words)

Provide a concise summary of the results. Comment on the validation status ("Pass"/"Fail") for each summary metric in Table T2. Explicitly state whether individual OS time-points (Table T1) and ECOG categories (Table T3) meet the ±5% concordance threshold. Conclude with an overall judgment on the simulation's external validity based on these benchmarks.

Formatting Rules

- Produce **Markdown tables only**; no plots, code, or images.
- Format percentages to **one decimal place**.
- Format Pearson r to **three decimal places**.
- Keep the interpretation paragraph strictly **within the 120-word limit**.

[S33.TRL.13.P30.LOG.csv] (Shown) OR [S37.TRL.14.P30.LOG.csv] OR [S40.TRL.15.P30.LOG.csv]

Table 9: Ref: S35b.VER.03.P34, S38b.VER.02.P35, S41b.VER.02.P36

Trial Charts: Prompt 33

You have access to the full simulated PDAC Cancer 100,000-patient Phase III clinical trial log file with the following columns: patient_id, arm, archetype, age, stage_iv, ecog, kras_g12c, gbrca, ca19_9, time_to_progression_or_death, time_to_death, and time_to_first_G3_AE. Generate the following 30 visualizations as separate PNG files in one folder, ensuring each plot is clearly titled and labeled. The control arm is Arm E.

List of 30 Visualizations:

1. Bar chart of patient counts per treatment arm, to confirm balanced randomization across all five arms.
2. Overlaid density plots of patient age distribution for each treatment arm, to visualize and compare the age profile across cohorts.
3. Stacked bar chart showing the distribution of ECOG performance status (0, 1, and 2) across all treatment arms, to verify baseline functional status balance.
4. Grouped bar chart comparing the percentage of patients with KRAS mutation status (kras_g12c) for each treatment arm.
5. Box plot of baseline CA 19-9 tumor marker levels by treatment arm, to assess the distribution and balance of this key prognostic biomarker.
6. Kaplan-Meier plot for Overall Survival (OS), comparing all five treatment arms on a single graph.
7. Kaplan-Meier plot for Progression-Free Survival (PFS), comparing all five treatment arms on a single graph.
8. Bar chart displaying the median Overall Survival (in months) for each arm, with error bars representing the 95% confidence interval.
9. Bar chart displaying the median Progression-Free Survival (in months) for each arm, with error bars representing the 95% confidence interval.
10. Bar chart of the 12-month Overall Survival rate for each treatment arm, to visually represent this key timepoint metric.
11. Kaplan-Meier plot for Time to First Grade ≥3 Adverse Event, comparing all treatment arms to visualize safety profiles over time.
12. Bar chart showing the overall incidence rate (%) of patients experiencing a Grade ≥3 Adverse Event within 24 months, for each treatment arm.
13. Forest plot or bar chart visualizing the Overall Survival Hazard Ratios (and 95% CIs) for each experimental arm relative to the control arm.
14. Forest plot or bar chart visualizing the Progression-Free Survival Hazard Ratios (and 95% CIs) for each experimental arm relative to the control arm.
15. Scatter plot of Time to Progression vs. Overall Survival for all patients, colored by treatment arm, to show the correlation between endpoints.
16. Violin plot showing the distribution of Overall Survival time for each treatment arm, to compare the full range and density of survival outcomes.
17. Kaplan-Meier plot for Overall Survival stratified by ECOG status (ECOG 0 vs. ECOG 1-2) for the most effective arm (Arm A) versus the control arm (Arm E).
18. Kaplan-Meier plot for Overall Survival stratified by KRAS mutation status (kras_g12c), comparing outcomes within the most effective arm (Arm A).
19. Kaplan-Meier plot for Overall Survival stratified by gBRCA mutation status, comparing outcomes for all arms combined.
20. Scatter plot of baseline CA 19-9 levels versus Overall Survival time for all patients, colored by treatment arm to identify prognostic value.
21. Bar chart comparing median Overall Survival between younger (<65 years) and older (≥65 years) patient subgroups, faceted by treatment arm.
22. Heatmap showing the Pearson correlation matrix between continuous variables: age, CA 19-9, time to progression, time to death, and time to first G3 AE.
23. A risk-benefit bubble chart where the X-axis is median PFS, Y-axis is median OS, and the bubble size represents the Grade ≥3 AE rate for each arm.
24. Swarm plot showing individual patient survival times for each arm, providing a granular view of the outcome distribution and censoring.
25. Cumulative incidence plot for Grade ≥3 AEs, with death as a competing risk, comparing the triplet arm (Arm A) to the control arm (Arm E).
26. Box plots comparing Overall Survival across different patient archetype groups to explore this novel variable.
27. Scatter plot of Time to First Grade ≥3 AE versus Overall Survival time, colored by treatment arm, to investigate the relationship between early toxicity and efficacy.
28. Waterfall plot of individual patient survival times in the most effective arm (Arm A), ordered from shortest to longest survival.
29. Grouped bar chart comparing median Progression-Free Survival in patients with high vs. low baseline CA 19-9 (split by the median), for each arm.
30. Stacked bar chart showing the cause of PFS events (progression vs. death) for each treatment arm, if such data can be inferred from the time-to-event variables.

[S33.TRL.13.P30.LOG.csv] OR [S37.TRL.14.P30.LOG.csv] OR [S40.TRL.15.P30.LOG.csv]

Table 10: Ref: S36.VIS.01.P33, S39.VIS.01.P33, S42.VIS.01.P33

Trial vs. Trial: Prompt 37 (I/II)

Based on the three provided clinical trial simulation reports ("Trial 1", "Trial 2", "Trial 3"), you are to perform a cross-trial verification analysis. Your task is to generate five new comparison tables. For this task, you will **only** use the data contained within the tables of the three provided reports (Table 5-1, 6-1, 7-1, 8-1, and 8-2).

Each new table must be constructed according to the specific instructions below, including exact dimensions, row/column names, cell content, and a final consistency score. The goal is to rigorously assess the stability and consistency of the simulation's outputs across the three runs.

General Instructions for All Tables

1. **Data Extraction:** For each metric in a new table, you will locate the corresponding values from the equivalent tables in all three trial reports (Trial 1, Trial 2, Trial 3). This will give you a set of three numerical values for each data point.
2. **Cell Value Calculation:** For each cell in columns C1 through C5, you must calculate and display three statistics for the corresponding set of three values:
    - **Mean:** The arithmetic average of the three values.
    - **Range:** The difference between the maximum and minimum of the three values.
    - **Standard Deviation (SD):** The sample standard deviation of the three values.
    - **Format:** Present these as (Mean, Range, SD) and round to two decimal places, unless the original data has more precision (e.g., CA 19-9).
3. **Consistency Score Calculation (Final Column):** The final column of each table is a "Row Consistency Score" on a scale of 1.0 to 10.0 in 0.1 increments. This score measures the stability of a given metric across all arms and all three trials.
    - **Method:** For a given row, collect all 15 data points (5 arms x 3 trials). Calculate the overall Mean and overall Standard Deviation (SD) for this set of 15 values.
    - **Formula:** Score = 10.0 * (1 - (Overall SD / Overall Mean)).
    - **Rules:** If the Overall Mean is zero, the score is 10.0 (as SD will also be zero, indicating perfect consistency). Round the final score to one decimal place.
4. **Example Calculations:** Below each generated table, provide three detailed example calculations as specified in each table's instructions. Each example must clearly show the source values, the intermediate steps, and the final result for both the cell statistics and the consistency score.

Prompt for New Tables

1. Verification Table 1: Cross-Trial Consistency of Baseline Characteristics (from Table 5-1s)
Instructions: Generate a table that analyzes the consistency of baseline patient characteristics across the three trials.

- **Title:** Verification Table 1: Cross-Trial Consistency of Baseline Characteristics
- **Dimensions:** 8 Rows x 6 Columns
- **Row Names:**
    - R1: Age (years, mean)
    - R2: Stage IV (%)
    - R3: ECOG 0 (%)
    - R4: ECOG 1 (%)
    - R5: ECOG 2 (%)
    - R6: KRAS-mutant (%)
    - R7: gBRCA-mutant (%)
    - R8: CA 19-9 (U/mL, mean)
- **Column Names:**
    - C1: Arm A (Mean, Range, SD)
    - C2: Arm B (Mean, Range, SD)
    - C3: Arm C (Mean, Range, SD)
    - C4: Arm D (Mean, Range, SD)
    - C5: Arm E (Mean, Range, SD)
    - C6: Row Consistency Score

Example Calculations to Provide Below Table 1:

1. **Cell (R1, C1):** Show the calculation for the Mean, Range, and SD for "Age (years, mean)" in Arm A.
2. **Cell (R4, C5):** Show the calculation for the Mean, Range, and SD for "ECOG 1 (%)" in Arm E.
3. **Score (R8, C6):** Show the calculation for the "Row Consistency Score" for the "CA 19-9" metric, including the collection of the 15 source values and the application of the scoring formula.

2. Verification Table 2: Cross-Trial Consistency of Primary Efficacy Outcomes (from Table 6-1s)
Instructions: Generate a table that analyzes the consistency of the primary efficacy outcomes across the three trials.

- **Title:** Verification Table 2: Cross-Trial Consistency of Primary Efficacy Outcomes
- **Dimensions:** 5 Rows x 6 Columns
- **Row Names:**
    - R1: Median PFS (mo)
    - R2: Median OS (mo)
    - R3: 12-month OS Rate (%)
    - R4: PFS HR vs Control
    - R5: OS HR vs Control
- **Column Names:**
    - C1: Arm A (Mean, Range, SD)
    - C2: Arm B (Mean, Range, SD)
    - C3: Arm C (Mean, Range, SD)
    - C4: Arm D (Mean, Range, SD)
    - C5: Arm E (Mean, Range, SD)
    - C6: Row Consistency Score
- **Example Calculations to Provide Below Table 2:**
    1. **Cell (R2, C1):** Show the calculation for "Median OS (mo)" in Arm A.
    2. **Cell (R4, C2):** Show the calculation for "PFS HR vs Control" in Arm B.
    3. **Score (R3, C6):** Show the calculation for the "Row Consistency Score" for the "12-month OS Rate (%)" metric.

Table 11: Ref: S43.TST.01.P37, S44.TST.02.P37, S45.TST.03.P37, S46.TST.04.P37, S47.TST.05.P37

Trial vs. Trial: Prompt 37 (II/II)

3. Verification Table 3: Cross-Trial Consistency of Safety Outcomes (from Table 7-1s)
Instructions: Generate a table that analyzes the consistency of the summary safety outcome across the three trials.

- **Title:** Verification Table 3: Cross-Trial Consistency of Safety Outcomes
- **Dimensions:** 1 Row x 6 Columns
- **Row Names:**
  - R1: Patients with $\geq$G3 AE (%)
- **Column Names:**
  - C1: Arm A (Mean, Range, SD)
  - C2: Arm B (Mean, Range, SD)
  - C3: Arm C (Mean, Range, SD)
  - C4: Arm D (Mean, Range, SD)
  - C5: Arm E (Mean, Range, SD)
  - C6: Row Consistency Score
- **Example Calculations to Provide Below Table 3:**
  1. **Cell (R1, C1):** Show the calculation for "Patients with $\geq$G3 AE (%)" in Arm A.
  2. **Cell (R1, C5):** Show the calculation for "Patients with $\geq$G3 AE (%)" in Arm E.
  3. **Score (R1, C6):** Show the calculation for the "Row Consistency Score" for the "Patients with $\geq$G3 AE (%)" metric.

4. Verification Table 4: Cross-Trial Consistency of Median PFS by Archetype (from Table 8-1s)
Instructions: Generate a table that analyzes the consistency of the median Progression-Free Survival (PFS) within each patient archetype across the three trials.

- **Title:** Verification Table 4: Cross-Trial Consistency of Median PFS by Archetype
- **Dimensions:** 7 Rows x 6 Columns
- **Row Names:**
  - R1: ARCH-01 (Young_Fit_Metastatic)
  - R2: ARCH-02 (Elderly_Frail_Metastatic)
  - R3: ARCH-03 (LAPC_Standard_Fitness)
  - R4: ARCH-04 (Young_Fit_BRCAm)
  - R5: ARCH-05 (Metastatic_KRAS_G12C)
  - R6: ARCH-06 (Metastatic_High_Stroma)
  - R7: ARCH-07 (Advanced_Refractory_PS1)
- **Column Names:**
  - C1: Arm A (Mean, Range, SD)
  - C2**:** Arm B (Mean, Range, SD)
  - C3: Arm C (Mean, Range, SD)
  - C4: Arm D (Mean, Range, SD)
  - C5: Arm E (Mean, Range, SD)
  - C6: Row Consistency Score
- **Example Calculations to Provide Below Table 4:**
  1. **Cell (R3, C1):** Show the calculation for Median PFS for "ARCH-03" in Arm A.
  2. **Cell (R5, C2):** Show the calculation for Median PFS for "ARCH-05" in Arm B.
  3. **Score (R2, C6):** Show the calculation for the "Row Consistency Score" for the "ARCH-02" metric.

5. Verification Table 5: Cross-Trial Consistency of Median OS by Archetype (from Table 8-2s)
Instructions: Generate a table that analyzes the consistency of the median Overall Survival (OS) within each patient archetype across the three trials.

- **Title:** Verification Table 5: Cross-Trial Consistency of Median OS by Archetype
- **Dimensions:** 7 Rows x 6 Columns
- **Row Names:**
  - R1: ARCH-01 (Young_Fit_Metastatic)
  - R2: ARCH-02 (Elderly_Frail_Metastatic)
  - R3: ARCH-03 (LAPC_Standard_Fitness)
  - R4: ARCH-04 (Young_Fit_BRCAm)
  - R5: ARCH-05 (Metastatic_KRAS_G12C)
  - R6: ARCH-06 (Metastatic_High_Stroma)
  - R7: ARCH-07 (Advanced_Refractory_PS1)
- **Column Names:**
  - C1: Arm A (Mean, Range, SD)
  - C2: Arm B (Mean, Range, SD)
  - C3: Arm C (Mean, Range, SD)
  - C4: Arm D (Mean, Range, SD)
  - C5: Arm E (Mean, Range, SD)
  - C6: Row Consistency Score
- **Example Calculations to Provide Below Table 5:**
  1. **Cell (R1, C4):** Show the calculation for Median OS for "ARCH-01" in Arm D.
  2. **Cell (R5, C1):** Show the calculation for Median OS for "ARCH-05" in Arm A.
  3. **Score (R7, C6):** Show the calculation for the "Row Consistency Score" for the "ARCH-07" metric.

[S43.DAT.02.TAB]

Table 12: Ref: S43.TST.01.P37, S44.TST.02.P37, S45.TST.03.P37, S46.TST.04.P37, S47.TST.05.P37

Model vs. Model Charts: Prompt 38

Prompt for Cross-Model Verification Analysis Visualizations

You have been provided with 5 verification analysis outputs from different AI models (grk4, grk3, ops4, g25p, o3pr) that were all given the same prompt template to analyze three trials for consistency.

Analysis Summary: Provide a two-paragraph explanation of findings regarding the correspondence between the AI models' outputs. Focus on: patterns of agreement/disagreement between models, specific metrics where models showed highest/lowest correspondence, systematic differences in calculation approaches, and implications for AI model reliability in clinical data analysis. Cite visualizations 01-10 throughout the analysis summary.

Generate 10 separate visualizations in Python scripts (numbered 01-10) as follows:

01. Heatmap showing Row Consistency Scores across all models (5 models x 28 total metrics from all tables)
02. Grouped bar chart comparing Mean calculations for Baseline Characteristics (Table 1) across all 5 models and all 5 arms
03. Scatter plot matrix showing pairwise model agreement for all Row Consistency Scores, with correlation coefficients
04. Box plot displaying the distribution of Standard Deviation calculations across models for Primary Efficacy Outcomes (Table 2)
05. Radar chart comparing each model's Row Consistency Scores for the 7 archetypes in Table 4 (Median PFS)
06. Line graph showing Range calculations across models for Safety Outcomes data, with error bars indicating inter-model variance
07. Parallel coordinates plot displaying how each model calculated statistics for CA 19-9 baseline values across all arms
08. Stacked bar chart showing the frequency of exact agreement vs. minor/major discrepancies between model pairs
09. Bubble chart plotting Mean vs. SD calculations for Median OS by Archetype (Table 5), with bubble size representing Range and color representing model
10. Diverging bar chart highlighting the largest positive and negative deviations from the median Row Consistency Score for each metric across all models

"Begin grk4 = Grok 4" "End grk4 = Grok 4" "Begin grk3 = Grok 3 Think" "End grk3 = Grok 3 Think" "Begin ops4 = Opus 4 Extended" "End ops4 = Opus 4 Extended" "Begin g25p = Gemini 2.5 Pro" "End g25p = Gemini 2.5 Pro" "Begin o3pr = o3-pro" "End o3pr = o3-pro" [S43.TST.01.P37] [S44.TST.02.P37] [S45.TST.03.P37] [S46.TST.04.P37] [S47.TST.05.P37]

Table 13: Ref: S48.VIS.01.P38

Trial vs. Trial Charts: Prompt 39

Prompt for Cross-Trial Reproducibility Synthesis Analysis

You have been provided with 5 verification analysis outputs from different AI models (grk4, grk3, ops4, g25p, o3pr) that independently analyzed the reproducibility of three clinical trials. Each model calculated consistency metrics across baseline characteristics, efficacy outcomes, safety data, and archetype-specific results.

Analysis Summary: Provide a two-paragraph explanation synthesizing the collective findings regarding the reproducibility of the three trials. Focus on: the overall reproducibility patterns identified across all five models, specific trial parameters showing highest/lowest consistency, biological vs. technical sources of variation, and implications for the simulation engine's reliability. Include statistical measures (mean consistency scores, median values, standard deviations, and Pearson's r correlations between trial parameters where applicable). Focus less on direct comparisons between the 5 analyses. Cite visualizations 01-10 throughout the analysis summary.

Generate 10 separate visualizations in Python scripts (numbered 01-10) as follows:

1. Heatmap showing the consensus Row Consistency Scores (averaged across all 5 models) for all 28 metrics, organized by table category (Baseline, Efficacy, Safety, Archetype PFS, Archetype OS)
2. Box plot displaying the distribution of consistency scores by metric category (Baseline vs. Efficacy vs. Safety vs. Archetype-specific), showing trial reproducibility patterns
3. Scatter plot with regression line showing the relationship between baseline characteristic consistency and primary efficacy outcome consistency across all metrics
4. Grouped bar chart comparing consistency scores for each treatment arm (A-E) across all metric categories, revealing arm-specific reproducibility patterns
5. Line graph showing how consistency scores vary by archetype (ARCH-01 through ARCH-07) for both PFS and OS outcomes, with confidence intervals
6. Correlation matrix heatmap showing Pearson's r values between different metric categories' consistency scores
7. Violin plot comparing the distribution of Mean, Range, and SD values across the three trials for key efficacy metrics
8. Parallel coordinates plot showing the trajectory of consistency scores from baseline → efficacy → safety → archetype outcomes for each treatment arm
9. Bubble chart plotting metric variance (y-axis) vs. clinical importance weight (x-axis), with bubble size representing consensus consistency score and color representing metric category
10. Waterfall chart showing the cumulative impact of each metric category on overall trial reproducibility, starting from perfect consistency (10.0) and showing decrements

"Begin grk4 = Grok 4" "End grk4 = Grok 4" "Begin grk3 = Grok 3 Think" "End grk3 = Grok 3 Think" "Begin ops4 = Opus 4 Extended" "End ops4 = Opus 4 Extended" "Begin g25p = Gemini 2.5 Pro" "End g25p = Gemini 2.5 Pro" "Begin o3pr = o3-pro" "End o3pr = o3-pro" [S43.TST.01.P37] [S44.TST.02.P37] [S45.TST.03.P37] [S46.TST.04.P37] [S47.TST.05.P37]

Table 14: Ref: S49.VIS.02.P39

Log vs. Report: Prompt 40 (I/II)

You are tasked with a meta-verification analysis. Using the provided data from "Trial 1," "Trial 2," and "Trial 3," you will generate six new comparison tables. The goal is to re-evaluate the consistency of discrepancies between reported and calculated data across the three trials using a revised methodology that corrects for issues in a previous analysis.

This new methodology introduces a more robust, context-aware scoring system to accurately assess consistency. It distinguishes between standard metrics and percentage-based metrics, applying a unique formula to each to prevent misinterpretation of consistency for high-magnitude percentage values. It also includes explicit rules for data parsing to handle non-numeric characters.

For this task, you will only use the data from the Discrepancy, Deviation, or Difference columns of the provided source tables (Tables 1-6 for each of the three trials).

General Instructions for All Tables

1. Data Pre-processing and Extraction:

- For each required data point, locate the corresponding value in the "Discrepancy," "Deviation," or "Difference" column from the equivalent source table (e.g., Table 2, "Mean Age (years) Deviation") in all three trials.
- Crucially, you must parse **only the numerical value** from each cell. Ignore all non-numeric text, symbols, and formatting.
    - **Examples:**
        - +0.3 mo should be parsed as 0.3.
        - –0.5% or -0.5% should be parsed as -0.5.
        - 0.2 years should be parsed as 0.2.
        - 86.1%[11†] should be parsed as 86.1.
        - A value of 0.0 or –0.0 should be parsed as 0.0.

2. Cell Value Calculation:

- For each cell in columns C1 through C5 (where applicable), you will calculate and display three statistics for the set of three parsed numerical values from the trials:
    - **Mean:** The arithmetic average of the three values.
    - **Range:** The difference between the maximum and minimum of the three values.
    - **Standard Deviation (SD):** The sample standard deviation of the three values.
- **Format:** Present these as (Mean, Range, SD) and round each statistic to two decimal places.

3. Row Consistency Score Calculation (Final Column):

The final column of each table is a "Row Consistency Score" on a scale of 1.0 to 10.0. This score measures the stability of the discrepancy for a given metric across the trials.

- **Method:** For a given row (metric), collect all underlying parsed numerical values (e.g., 5 arms x 3 trials = 15 values, or 1 arm x 3 trials = 3 values for Table 2). Calculate the **Overall Mean** and **Overall Standard Deviation (SD)** for this set of values.
- **Select the appropriate formula based on the metric type:**
    - **Formula A (Standard Metrics):** Use for Tables 1, 3, and 4 (Patient Counts, Months).
      Consistency Score = 10.0 * (1 - (Overall SD / (|Overall Mean| + 1.0)))
        - **Rationale:** The addition of 1.0 to the denominator stabilizes the formula, preventing the score from becoming artificially low when the Overall Mean of the discrepancies is close to zero.
    - **Formula B (Percentage-Based Metrics):** Use for Tables 2, 5, and 6 (All metrics ending in "%").
      Consistency Score = 10.0 * (1 - (Overall SD / (|Overall Mean| + 10.0)))
        - **Rationale:** For percentage-based data, absolute differences are often small (e.g., +/- 1-2%). The standard formula can incorrectly penalize tight clustering of these small values. The larger + 10.0 scaling factor makes the score robust to this effect by evaluating the standard deviation of the discrepancies relative to a larger denominator. This better reflects high consistency when small discrepancies are tightly grouped around a mean close to zero.
- **Rules for Both Formulas:**
    - If the Overall SD is zero (indicating perfect consistency), the score is **10.0**.
    - Round the final score to **one decimal place**.
    - If the calculated score is less than 1.0, it must be reported as **1.0**. The maximum score is **10.0**.

4. Example Calculations:

- Below each generated table, provide three detailed example calculations as specified in that table's instructions. Each example must clearly show:
    - The source values from the three trials (after parsing).
    - The intermediate steps and final result for the cell statistics (Mean, Range, SD).
    - The intermediate steps and final result for the Row Consistency Score, **explicitly stating whether Formula A or Formula B was used**.

Instructions for New Tables

1. Meta-Verification Table 1: Cross-Verification Consistency of Cohort Distribution Discrepancy

- **Title:** Meta-Verification Table 1: Cross-Verification Consistency of Cohort Distribution Discrepancy
- **Dimensions:** 1 Row x 6 Columns
- **Row Name:** R1: Patient Count Discrepancy
- **Column Names:** C1: Arm A (Mean, Range, SD), C2: Arm B (Mean, Range, SD), C3: Arm C (Mean, Range, SD), C4: Arm D (Mean, Range, SD), C5: Arm E (Mean, Range, SD), C6: Row Consistency Score
- **Special Instruction for Score Calculation:** The Row Consistency Score must be calculated using **Formula A (Standard Metrics)**.
- **Example Calculations:** Show the calculations for Cell (R1, C1), Cell (R1, C4), and the Score for (R1, C6).

Table 15: Ref: S50.TST.01.P40, S51.TST.02.P40, S52.TST.03.P40, S53.TST.04.P40, S54.TST.05.P40

2. Meta-Verification Table 2: Cross-Verification Consistency of Baseline Characteristic Deviations (Arm A)
- **Title:** Meta-Verification Table 2: Cross-Verification Consistency of Baseline Characteristic Deviations (Arm A)
- **Dimensions:** 5 Rows x 2 Columns
- **Row Names:** R1: Mean Age (years) Deviation, R2: Stage IV (%) Deviation, R3: ECOG 1 (%) Deviation, R4: KRAS-mutant (%) Deviation, R5: gBRCA-mutant (%) Deviation
- **Column Names:** C1: Arm A (Mean, Range, SD), C2: Row Consistency Score
- **Special Instructions for Score Calculation:**
  - The Row Consistency Score for each row must be calculated using only the 3 underlying values from Arm A (1 arm x 3 trials).
  - For this table, the Row Consistency Score must be calculated using **Formula B (Percentage-Based Metrics)** for all rows, as they are all percentages (even if the unit isn't in the title).
- **Example Calculations:** Show the calculations for Cell (R1, C1), the Score for (R2, C2), and the Score for (R4, C2).

3. Meta-Verification Table 3: Cross-Verification Consistency of Median OS Difference
- **Title:** Meta-Verification Table 3: Cross-Verification Consistency of Median OS Difference
- **Dimensions:** 1 Row x 6 Columns
- **Row Name:** R1: Median OS Difference (months)
- **Column Names:** C1: Arm A (Mean, Range, SD), C2: Arm B (Mean, Range, SD), C3: Arm C (Mean, Range, SD), C4: Arm D (Mean, Range, SD), C5: Arm E (Mean, Range, SD), C6: Row Consistency Score
- **Special Instruction for Score Calculation:** The Row Consistency Score must be calculated using **Formula A (Standard Metrics)**.
- **Example Calculations:** Show the calculations for Cell (R1, C1), Cell (R1, C5), and the Score for (R1, C6).

4. Meta-Verification Table 4: Cross-Verification Consistency of Median PFS Difference
- **Title:** Meta-Verification Table 4: Cross-Verification Consistency of Median PFS Difference
- **Dimensions:** 1 Row x 6 Columns
- **Row Name:** R1: Median PFS Difference (months)
- **Column Names:** C1: Arm A (Mean, Range, SD), C2: Arm B (Mean, Range, SD), C3: Arm C (Mean, Range, SD), C4: Arm D (Mean, Range, SD), C5: Arm E (Mean, Range, SD), C6: Row Consistency Score
- **Special Instruction for Score Calculation:** The Row Consistency Score must be calculated using **Formula A (Standard Metrics)**.
- **Example Calculations:** Show the calculations for Cell (R1, C2), Cell (R1, C5), and the Score for (R1, C6).

5. Meta-Verification Table 5: Cross-Verification Consistency of 12-Month OS Rate Difference
- **Title:** Meta-Verification Table 5: Cross-Verification Consistency of 12-Month OS Rate Difference
- **Dimensions:** 1 Row x 6 Columns
- **Row Name:** R1: 12-Month OS Rate Difference (%)
- **Column Names:** C1: Arm A (Mean, Range, SD), C2: Arm B (Mean, Range, SD), C3: Arm C (Mean, Range, SD), C4: Arm D (Mean, Range, SD), C5: Arm E (Mean, Range, SD), C6: Row Consistency Score
- **Special Instruction for Score Calculation:** The Row Consistency Score must be calculated using **Formula B (Percentage-Based Metrics)**.
- **Example Calculations:** Show the calculations for Cell (R1, C1), Cell (R1, C3), and the Score for (R1, C6).

6. Meta-Verification Table 6: Cross-Verification Consistency of ≥G3 AE Rate Difference
- **Title:** Meta-Verification Table 6: Cross-Verification Consistency of ≥G3 AE Rate Difference
- **Dimensions:** 1 Row x 6 Columns
- **Row Name:** R1: ≥G3 AE Rate Difference (%)
- **Column Names:** C1: Arm A (Mean, Range, SD), C2: Arm B (Mean, Range, SD), C3: Arm C (Mean, Range, SD), C4: Arm D (Mean, Range, SD), C5: Arm E (Mean, Range, SD), C6: Row Consistency Score
- **Special Instruction for Score Calculation:** The Row Consistency Score must be calculated using **Formula B (Percentage-Based Metrics)**.
- **Example Calculations:** Show the calculations for Cell (R1, C2), Cell (R1, C4), and the Score for (R1, C6).

[S44.DAT.03.TAB]

Table 16: Ref: S50.TST.01.P40, S51.TST.02.P40, S52.TST.03.P40, S53.TST.04.P40, S54.TST.05.P40

Log vs. Report vs. Model Charts: Prompt 41

Prompt for Cross-Model Meta-Verification Analysis Visualizations

You have been provided with 5 verification analysis outputs from different AI models (use these terms grk4, grk3, ops4, g25p, o3pr) that were all given the same prompt template to analyze three clinical trials for meta-verification consistency.

Analysis Summary: Provide a two-paragraph explanation of findings regarding the correspondence between the AI models' outputs. Focus on: patterns of agreement/disagreement between models in their meta-verification calculations, specific tables where models showed highest/lowest correspondence in Row Consistency Scores, systematic differences in statistical calculation approaches (mean, range, SD), and implications for AI model reliability in meta-analysis of clinical trial discrepancies. Cite visualizations 01-10 throughout the analysis summary.

Generate 10 separate visualizations in Python scripts (numbered 01-10) as follows:

1. Heatmap showing Row Consistency Scores across all models (5 models x 6 tables) with annotations for exact values and color gradient from 8.0 to 10.0
2. Grouped bar chart comparing Mean calculations for Table 2 (Baseline Characteristic Deviations) across all 5 models for each of the 5 baseline characteristics
3. Scatter plot matrix showing pairwise model agreement for all Row Consistency Scores across the 6 tables, with correlation coefficients and regression lines
4. Box plot displaying the distribution of Standard Deviation calculations across models for Table 3 (Median OS Difference) for all 5 arms
5. Radar chart comparing each model's Row Consistency Scores for all 6 meta-verification tables, with separate traces for each model
6. Line graph showing Range calculations across models for Table 6 (≥G3 AE Rate Difference) for all 5 arms, with confidence intervals
7. Parallel coordinates plot displaying how each model calculated statistics (Mean, Range, SD) for Table 4 (Median PFS Difference) across all arms
8. Stacked bar chart showing the frequency of exact agreement (difference < 0.01), minor discrepancies (0.01-0.1), and major discrepancies (>0.1) between model pairs for all calculated values
9. Bubble chart plotting Mean vs. SD calculations for Table 5 (12-Month OS Rate Difference) by arm, with bubble size representing Range and color representing model
10. Diverging bar chart highlighting the largest positive and negative deviations from the median Row Consistency Score for each table across all models, sorted by magnitude of deviation

"Begin grk4 = Grok 4" "End grk4 = Grok 4" "Begin grk3 = Grok 3 Think" "End grk3 = Grok 3 Think" "Begin ops4 = Opus 4 Extended" "End ops4 = Opus 4 Extended" "Begin g25p = Gemini 2.5 Pro" "End g25p = Gemini 2.5 Pro" "Begin o3pr = ChatGPT o3-pro" "End o3pr = ChatGPT o3-pro" [S50.TST.01.P40] [S51.TST.02.P40] [S52.TST.03.P40] [S53.TST.04.P40] [S54.TST.05.P40]

Table 17: Ref: S55.VIS.01.P41

Log vs. Report vs. Trial Charts: Prompt 42

Prompt for Meta-Verification Cross-Trial Consistency Analysis

You have been provided with 5 meta-verification analysis outputs from different AI models (grk4, grk3, ops4, g25p, o3pr) that independently analyzed the consistency of discrepancies, deviations, and differences across three clinical trials. Each model calculated meta-verification consistency scores using standardized formulas across six key comparison dimensions: cohort distribution, baseline characteristics, median OS differences, median PFS differences, 12-month OS rate differences, and grade ≥3 adverse event rate differences.

Analysis Summary: Provide a two-paragraph explanation synthesizing the collective findings regarding the meta-verification consistency patterns identified across all five models. Focus on: the overall consistency patterns in measurement discrepancies across trials, specific meta-verification tables showing highest/lowest row consistency scores, the relationship between baseline deviation consistency and outcome difference consistency, and implications for understanding systematic vs. random sources of variation in trial reporting. Include statistical measures (mean row consistency scores ranging from 8.8-10.0, coefficient of variation across models, inter-model agreement metrics, and Spearman's rank correlations between different meta-verification dimensions where applicable). Focus less on direct comparisons between the 5 analyses. Cite visualizations 01-10 throughout the analysis summary.

Generate 10 separate visualizations in Python scripts (numbered 01-10) as follows:

1. Heatmap showing the consensus Row Consistency Scores (averaged across all 5 models) for all 6 meta-verification tables, with cells color-coded by score magnitude (8.0-10.0 scale) and annotated with inter-model standard deviations
2. Box plot displaying the distribution of row consistency scores by meta-verification category (Cohort vs. Baseline vs. OS Difference vs. PFS Difference vs. 12-Month OS Rate vs. AE Rate), revealing patterns in measurement consistency
3. Scatter plot matrix showing pairwise relationships between consistency scores from different meta-verification tables, with regression lines and R^2 values for each pair
4. Radar chart comparing the consistency profile of each treatment arm (A-E) across all meta-verification dimensions, showing arm-specific measurement reliability patterns
5. Line graph with error bars showing how mean cell statistics (Mean, Range, SD) vary across treatment arms for each meta-verification table, with separate panels for each table
6. Clustered heatmap showing the correlation structure between all row consistency scores and their underlying cell statistics (means, ranges, SDs), with dendrogram showing hierarchical relationships
7. Violin plot comparing the distribution of consistency scores between Formula A metrics (standard) vs. Formula B metrics (percentage-based), overlaid with individual data points
8. 3D surface plot showing the relationship between overall mean values, overall SD values, and resulting consistency scores across all metrics, illustrating the scoring formula landscapes
9. Sankey diagram showing the flow from raw trial discrepancy values through cell statistics (Mean, Range, SD) to final row consistency scores for each meta-verification table
10. Ridge plot (joy plot) showing the distribution of individual trial values that contribute to each meta-verification table's consistency scores, stacked by table type and colored by consistency score magnitude

"Begin grk4 = Grok 4" "End grk4 = Grok 4" "Begin grk3 = Grok 3 Think" "End grk3 = Grok 3 Think" "Begin ops4 = Opus 4 Extended" "End ops4 = Opus 4 Extended" "Begin g25p = Gemini 2.5 Pro" "End g25p = Gemini 2.5 Pro" "Begin o3pr = ChatGPT o3-pro" "End o3pr = ChatGPT o3-pro" [S50.TST.01.P40] [S51.TST.02.P40] [S52.TST.03.P40] [S53.TST.04.P40] [S54.TST.05.P40]

Table 18: Ref: S56.VIS.02.P42

In Silico Trial Overview: Prompt 43

"Instructions Start"
Analyze, utilize, and cite the provided documents to produce a comprehensive virtual study overview of the completed 100,000-patient virtual triplicate simulations. Produce a single, detailed report in the "Executive Summary", "Technical Details", "Key Insights" format. Use large, interpretable markdown tables designated with appropriate rows R1, R2.. and columns C1, C2.. suitable for downstream data extraction and visualization.

Input Files for Processing:

- Trial reports, log file verifications, external validations, and visualizations for a 100,000-patient, 5-arm in-silico Phase III simulation run in triplicate and verified by multiple AI models (grk4, grk3, ops4, g25p, o3pr). Log file verfications correspond to files such as S35.VER.02.P32.

A. Virtual Study Triplicate Details

1. **Four Tables with specific rows R1, R2.. and columns C1, C2..**
2. **Fill in Details of each cell with combined data from the included files below**
3. **Table 01: 3 Virtual Trials - Provide Additional Details**
    - **Study Title/Identifier**
    - **Primary Goal**
    - **Trial Phase Equivalence** (Phase III details)
    - **Study Design** (5-arm in-silico simulation)
    - **Trial Arms** (List the specific arms for each)
    - **Patient Population Size**
    - **Patient Archetypes** (7 archetypes)
4. **Table 02: 3 Virtual Trial Details - Provide Additional Details**
    - **Drug Combination(s)** (Note the shared core triplet)
    - **Patient Data Granularity** (Describe the level of detail for virtual patient creation)
    - **Modeling Architecture** (100K trial's exponential survival model)
    - **Project Timeline**
    - **Primary Endpoints**
    - **Key AI Models Utilized** (List for both, based on the provided information)
5. **Table 03: Benefits and Drawbackss - Provide Additional Details**
    - **Itemized Benefits** of the Completed 100K Patient Triplicate Simulation: Pay particular attention to all benefits derived from the completed triplicate simulation. Analyze the value of its speed, scale, and robust cross-model/cross-trial verification (as seen in files such as S43, S48, S49, S50, S55, S56).
    - **Itemized Drawbacks** of the Completed 100K Patient Triplicate Simulation: List the drawbacks and limitations of the 100K trial's approach, considering factors like its simplified patient models and potential for "black-box" objections. Detail how methods used for the simulated trials could be improved in future studies.
6. **Table 04: Reproducibility Findings - Most Comprehensively Detailed Table**
    - **Validation** (100K trial's internal log verification and external validation against Flatiron data) Provide full detail regarding all results reported. Be sure to include inclusion of data synergies from Table T1 - OS concordance, Table T2 - OS Summary Metrics, and Table T3-ECOG Confidence scores.
    - **Reproducibility** (Overall reproducibility metrics of triplicate runs and cross-model verification in files like S43, S50, S55, S56). Include analysis of visualization scripts in Python from files like S48.VIS.01.P38. Be sure to include in full detail how reproducibility across the three simulated trials was observed or not observed by analyzing, utilizing, and citing specific documents included below.

"Instructions End"
"Use Model Abbreviations in Output Start"
grk4 = Grok 4, grk3 = Grok 3, ops4 = Opus 4, g25p = Gemini 2.5 Pro, o3pr = ChatGPT o3-pro, o3ph = ChatGPT o3-pro Research
"Use Model Abbreviations in Output End"
"File Descriptions Start"
S33.TRL.13.P30 = Trial 1 (Example)
S35.VER.02.P32 = Trial 1 tables vs. log file verifications (Example)
S35b.VER.03.P34 = Trial 1 external validation of log file (Example)
S36.VIS.01.P33 = Trial 1 visualizations (Example)
S37.TRL.14.P30 = Trial 2
S38.VER.01.P32 = Trial 2 tables vs. log file verifications
S38b.VER.02.P35 = Trial 2 external validation of log file
S39.VIS.01.P33 = Trial 2 visualizations
S40.TRL.15.P30 = Trial 3
S41.VER.01.P32 = Trial 3 tables vs. log file verifications
S41b.VER.02.P36 = Trial 3 external validation of log file
S42.VIS.01.P33 = Trial 3 visualizations
S43.TST.01.P37 = grk4 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output) (Example)
S44.TST.02.P37 = grk3 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output)
S45.TST.03.P37 = ops4 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output)
S46.TST.04.P37 = g25p 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output)
S47.TST.05.P37 = o3pr 3 Trial Tables Cross-Verification of Dataset 2 (cross-trial verification, 5 table output)
S48.VIS.01.P38 = 3 Trial Tables Cross-Verification: Visualize Models (Example)
S49.VIS.02.P39 = 3 Trial Tables Cross-Verification: Visualize Trials (Example)
S50.TST.01.P40 = grk4 Meta-Verification Tables Cross-Trial of Dataset 3
S35.VER.02.P32, S38.VER.01.P32, S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output) (Example)
S51.TST.02.P40 = grk3 Meta-Verification Tables Cross-Trial of Dataset 3
S35.VER.02.P32, S38.VER.01.P32, S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output)
S52.TST.03.P40 = ops4 Meta-Verification Tables Cross-Trial of Dataset 3
S35.VER.02.P32, S38.VER.01.P32, S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output)
S53.TST.04.P40 = g25p Meta-Verification Tables Cross-Trial of Dataset 3
S35.VER.02.P32, S38.VER.01.P32, S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output)
S54.TST.05.P40 = o3pr Meta-Verification Tables Cross-Trial of Dataset 3
S35.VER.02.P32, S38.VER.01.P32, S41.VER.01.P32 (3 report tables vs. log (S Files) vs. 3 trials. 6 table output)
S55.VIS.01.P41 = Verification Tables Cross-Trial: Visualize Models (Example)
S56.VIS.02.P42 = Verification Tables Cross-Trial: Visualize Trials (Example) "File Descriptions End"

[S33.TRL.13.P30] [S35.VER.02.P32] [S35b.VER.03.P34] [S36.VIS.01.P33] [S37.TRL.14.P30] [S38.VER.01.P32] [S38b.VER.02.P35] [S39.VIS.01.P33] [S40.TRL.15.P30] [S41.VER.01.P32] [S41b.VER.02.P36] [S42.VIS.01.P33] [S43.TST.01.P37] [S44.TST.02.P37] [S45.TST.03.P37] [S46.TST.04.P37] [S47.TST.05.P37] [S48.VIS.01.P38] [S49.VIS.02.P39] [S50.TST.01.P40] [S51.TST.02.P40] [S52.TST.03.P40] [S53.TST.04.P40] [S54.TST.05.P40] [S55.VIS.01.P41] [S56.VIS.02.P42]

Table 19: Ref: S57.REP.01.P43

Trial Overview Charts: Prompt 43b1

Based on the included comprehensive analysis of the 100,000-patient triplicate simulation study evaluating novel therapies for advanced Pancreatic Ductal Adenocarcinoma (PDAC), please generate 10 separate visualization scripts in Python that effectively communicate the key clinical findings, validation results, and methodological insights from this virtual trial. The visualizations should help stakeholders understand the efficacy-toxicity trade-offs, biomarker importance, and robustness of the simulation methodology.

Please create the following visualizations using Python:

01) Kaplan-Meier Survival Curves: Display overall survival curves for all 5 treatment arms with median OS values and confidence intervals annotated
02) Forest Plot of Hazard Ratios: Show OS and PFS hazard ratios with 95% CIs for Arms A-D versus control Arm E to visualize treatment effects
03) Stacked Bar Chart of Adverse Events: Compare Grade 3+ adverse event rates across all 5 arms highlighting the efficacy-toxicity trade-off
04) Heatmap of Archetype-Specific Outcomes: Display median OS across 7 patient archetypes and 5 treatment arms to identify subgroup benefits
05) Radar Chart of External Validation: Compare simulated control arm metrics against Flatiron real-world data for OS% at multiple timepoints and ECOG distribution
06) Box Plot of Cross-Trial Reproducibility: Show the distribution of key metrics across the three simulation runs demonstrating consistency
07) Waterfall Plot of KRAS G12C Response: Illustrate the differential treatment benefit for KRAS G12C patients across arms containing versus not containing Daraxonrasib
08) Scatter Plot Matrix of AI Model Agreement: Display pairwise correlations between the 5 AI models' consistency scores with clustering patterns
09) Sankey Diagram of Patient Flow: Visualize patient allocation across arms and progression through key clinical milestones including death and progression events
10) Combined Efficacy-Safety Bubble Plot: Plot median OS versus Grade 3+ AE rates for all arms with bubble size representing patient numbers to aid treatment selection decisions.
"Start Report" "End Report"

[S57.REP.01.P43]

Table 20: Ref: S57b.VIS.01.P43b

Trial Overview Charts: Prompt 43b2

Based on the included comprehensive analysis of the 100,000-patient triplicate simulation study evaluating novel therapies for advanced Pancreatic Ductal Adenocarcinoma (PDAC), please generate 10 separate visualization scripts in Python that effectively communicate the key clinical findings, validation results, and methodological insights from this virtual trial. The visualizations should help stakeholders understand the efficacy-toxicity trade-offs, biomarker importance, and robustness of the simulation methodology.

Please create the following visualizations using Python:

01) Kaplan-Meier Survival Curves: Display overall survival curves for all 5 treatment arms with median OS values and confidence intervals annotated
02) Forest Plot of Hazard Ratios: Show OS and PFS hazard ratios with 95% CIs for Arms A-D versus control Arm E to visualize treatment effects
03) Stacked Bar Chart of Adverse Events: Compare Grade 3+ adverse event rates across all 5 arms highlighting the efficacy-toxicity trade-off
04) Heatmap of Archetype-Specific Outcomes: Display median OS across 7 patient archetypes and 5 treatment arms to identify subgroup benefits
05) Radar Chart of External Validation: Compare simulated control arm metrics against Flatiron real-world data for OS% at multiple timepoints and ECOG distribution
06) Box Plot of Cross-Trial Reproducibility: Show the distribution of key metrics across the three simulation runs demonstrating consistency
07) Waterfall Plot of KRAS G12C Response: Illustrate the differential treatment benefit for KRAS G12C patients across arms containing versus not containing Daraxonrasib
08) Scatter Plot Matrix of AI Model Agreement: Display pairwise correlations between the 5 AI models' consistency scores with clustering patterns
09) Sankey Diagram of Patient Flow: Visualize patient allocation across arms and progression through key clinical milestones including death and progression events
10) Combined Efficacy-Safety Bubble Plot: Plot median OS versus Grade 3+ AE rates for all arms with bubble size representing patient numbers to aid treatment selection decisions.
"Start Report" "End Report"

[S57.REP.01.P43]

Table 21: Ref: S57b.VIS.01.P43b

Meta-Analysis: Prompt 44 (I/III)

Produce a complete, audit-ready "Comparative Clinical Metrics Meta-Analysis of the 100,000-Patient Virtual Trial Triplicate" as described in the provided report against other publicly available in-silico and real-world clinical trials in advanced Pancreatic Ductal Adenocarcinoma (PDAC) from 2010-2025. The primary focus of this analysis is a rigorous comparison of clinical trial metrics, designed to produce data and tables suitable for advanced downstream data visualization.

Use large, interpretable markdown tables designated with the strict R1, R2.. and C1, C2.. format for all tables. The primary data source for the 100,000-patient trial triplicate is exclusively the provided "Start Report" text. All quantitative data for external studies must be traceable via a direct URL. For any metric not explicitly stated in a source, state 'N/R' (Not Reported).

Return a single output containing the sections in this order:

Abstract (structured, $\leq$300 words)

- **Background:** Briefly state the challenges of traditional PDAC clinical trials and the emergence of in-silico trials as a tool for hypothesis generation and trial optimization.
- **Objective:** To conduct a systematic meta-analysis comparing the clinical efficacy, safety, and methodological parameters of the 100K-patient virtual trial (from the provided report) against other published in-silico and real-world interventional PDAC trials.
- **Methods:** Outline the data sources (provided report, PubMed, ClinicalTrials.gov), search strategy, study selection criteria (PRISMA), and the main data points for comparison (OS, PFS, AE rates, subgroup effects). Mention the development of a quantitative Efficacy-Toxicity Score (ETS) for head-to-head comparison.
- **Results:** Summarize the key comparative findings, including the relative performance of the virtual triplet arm (Arm A), the identification of concordance and discordance (e.g., ECOG mismatch), and the outcome of the head-to-head ETS scoring.
- **Conclusions:** State the main conclusions regarding the clinical utility and methodological standing of the 100K-patient simulation in the context of other PDAC research.
- **Registration:** PROSPERO Registration Number: [Placeholder]

Plain-language summary ($\leq$250 words)

Provide a clear, non-technical summary explaining what virtual clinical trials are, how the 100,000-patient simulation was compared to other computer-based and real-patient trials for pancreatic cancer, and what the main takeaways are for researchers designing future cancer studies.

Background

Briefly describe the high failure rates, long timelines, and significant costs associated with traditional oncology clinical trials, specifically in a challenging disease like PDAC. Introduce in-silico (computer-simulated) clinical trials as an emerging methodology to de-risk, accelerate, and optimize drug development. State that this meta-analysis will contextualize a large-scale virtual trial within the existing landscape of both virtual and real-world research.

Objectives

- The primary objective is to systematically compare the clinical trial metrics (efficacy, safety, patient characteristics, and outcomes) of the 100,000-patient virtual trial triplicate (as detailed in the provided report) against:
  1. Other publicly available in-silico PDAC trials.
  2. Pivotal real-world interventional Phase II and Phase III PDAC clinical trials. (Always prefer Phase III trials).
- The secondary objective is to develop and apply a quantitative scoring model to facilitate a direct head-to-head comparison of the therapeutic regimens across different study types and to identify key research gaps for future in-silico modeling.

Methods

- **Data Sources:** The primary data for the 100K-patient triplicate simulation will be extracted exclusively from the provided "Start Report" text. External data for comparator studies will be sourced from PubMed, ClinicalTrials.gov, ASCO/ESMO meeting abstracts, and peer-reviewed literature published between January 1, 2010, and December 31, 2025.
- **Search Strategy:** Specify the search terms used for external studies (e.g., "pancreatic adenocarcinoma," "PDAC," "in-silico," "virtual trial," "computational model," "Phase III," "Phase II," "Overall Survival"). State that the search is limited to English-language publications.
- **Study Selection:** Provide a PRISMA flow count in a table format.

R	C1: Stage	C2: Count
R1	Records identified from databases	[Number]
R2	Records removed before screening (e.g., duplicates)	[Number]
R3	Records screened	[Number]
R4	Records excluded	[Number]
R5	Reports sought for retrieval	[Number]
R6	Reports not retrieved	[Number]
R7	Reports assessed for eligibility	[Number]
R8	Reports excluded (with reasons)	[Number]
R9	Studies included in qualitative synthesis	[Number]
R10	Studies included in quantitative synthesis	[Number]

A. Virtual Study Comparison to Existing In-Silico PDAC Trials

Table 1: Comparative Clinical and Methodological Metrics of In-Silico PDAC Trials

Instructions: Populate this table by extracting data for C2 and C3 directly and exclusively from the provided "Start Report". Every cell for C2 and C3 must be filled; do not leave any as [Value from analysis of report]. For C4 and C5, find and cite credible published in-silico PDAC studies. **Prioritize selecting comparator studies that, at a minimum, report N, OS (or survival endpoint), and modeling architecture to ensure a meaningful comparison.** If a metric is not reported (N/R) in the external study, state that clearly.

Table 22: Ref: S58.REP.02.P44

Meta-Analysis: Prompt 44 (II/III)

R	C1: Metric / Parameter	C2: 100K Triplicate (Control Arm E)	C3: 100K Triplicate (Triplet Arm A)	C4: Comparator In-Silico Study 1	C5: Comparator In-Silico Study 2
R1	Patient Population Size (N)	20,000	20,000	[Value]	[Value]
R2	Patient Profile Summary	Fitter profile; ECOG 0/1/2 mismatch vs. RWD	Fitter profile; ECOG 0/1/2 mismatch vs. RWD	[Brief Description]	[Brief Description]
R3	Modeling Architecture	Exponential survival model (Weibull k=1.0)	Exponential survival model (Weibull k=1.0)	[e.g., Agent-Based, QSP, PK/PD]	[e.g., Agent-Based, QSP, PK/PD]
R4	Median Overall Survival (OS)	6.1 months	8.7 months	[Value]	[Value]
R5	OS Hazard Ratio (HR vs. Control)	1.00 (Reference)	$\sim$0.69	[Value or N/R]	[Value or N/R]
R6	Median Progression-Free Survival (PFS)	3.1 months	N/R	[Value or N/R]	[Value or N/R]
R7	PFS Hazard Ratio (HR vs. Control)	1.00 (Reference)	N/R	[Value or N/R]	[Value or N/R]
R8	Grade $\geq$3 Adverse Events (%)	76.5%	94.0%	[Value or N/R]	[Value or N/R]
R9	Defined Patient Archetypes	7 Archetypes (ARCH-01 to ARCH-07)	7 Archetypes (ARCH-01 to ARCH-07)	[List or describe, or N/R]	[List or describe, or N/R]
R10	Key Subgroup Finding	N/A (Control)	Enhanced benefit in ARCH-05 (KRAS G12C)	[Describe key finding or N/R]	[Describe key finding or N/R]
R11	Source (URL / Report)	Source: Report	Source: Report	[URL to publication]	[URL to publication]

B. Virtual Study Comparison to Real-World In-Person PDAC Trials

Table 2: Comparative Clinical Metrics of Virtual vs. Real-World PDAC Trials

Instructions: Populate this table using the report for C2 and C3. For C4, C5, and C6, use data from well-known, pivotal Phase III and Phase II PDAC trials (e.g., MPACT, NAPOLI-1, PRODIGE 24) and provide URLs. For any virtual arm metric not directly stated in the report (e.g., Arm D OS), calculate it if a clear basis (e.g., HR and baseline) is provided. State that the value is Calculated. If no basis exists (e.g., AE% for Arm D), state N/R.

R	C1: Metric / Parameter	C2: 100K Triplicate (Triplet Arm A)	C3: 100K Triplicate (Doublet Arm D)	C4: Real-World Phase III (e.g., MPACT)	C5: Real-World Phase III (e.g., NAPOLI-1)	C6: Real-World Phase II or III (Specify)
R1	Study / Regimen	Triplet (Dara+Mita+nal-IRI)	Doublet (Dara+Mita)	Gemcitabine + nab-Paclitaxel	nal-IRI + 5-FU/LV	[Regimen Name]
R2	Patient Population Size (N)	20,000	20,000	[Value, e.g., 861]	[Value, e.g., 417]	[Value]
R3	Baseline ECOG PS 0/1 (%)	>95% (Failed validation vs. RWD)	>95% (Failed validation vs. RWD)	[Value]	[Value]	[Value]
R4	Median Overall Survival (OS)	8.7 months	[Calculated Value from HR $\sim$0.76]	[Value, e.g., 8.5 mo]	[Value, e.g., 6.1 mo]	[Value]
R5	OS Hazard Ratio (HR vs. SoC)	$\sim$0.69	$\sim$0.76	[Value, e.g., 0.72]	[Value, e.g., 0.67]	[Value]
R6	Median Progression-Free Survival	N/R	N/R	[Value, e.g., 5.5 mo]	[Value, e.g., 3.1 mo]	[Value]
R7	PFS Hazard Ratio (HR vs. SoC)	N/R	N/R	[Value, e.g., 0.69]	[Value, e.g., 0.56]	[Value]
R8	Grade $\geq$3 Adverse Events (%)	94.0%	N/R	[Value, e.g., 84%]	[Value, e.g., 79%]	[Value]
R9	Objective Response Rate (ORR) (%)	N/R	N/R	[Value, e.g., 23%]	[Value, e.g., 16%]	[Value]
R10	Source (URL / Report)	Source: Report	Source: Report	[URL to publication]	[URL to publication]	[URL to publication]

C. Quantitative Head-to-Head Comparison and Pooled Analysis

Instructions: Create a comprehensive "flat" table suitable for data processing and visualization. Pool the key metrics from all selected studies (virtual and real-world) into this single table. **Ensure each experimental arm has its corresponding control arm listed in the table to provide the baseline for delta (Δ) calculations.** Then, calculate the Efficacy-Toxicity Score (ETS) for each experimental arm as defined below.

- **Pooling and Scoring Instructions:**
  - For each study, identify the experimental arm(s) and its corresponding control arm. The control arm data is used for calculating the benefit and score.
  - Calculate the **Efficacy-Toxicity Score (ETS)**: The ETS provides a single value to compare the overall clinical utility of a regimen, balancing its survival benefit against its toxicity burden, relative to its own control.

Table 23: Ref: S58.REP.02.P44

- **ETS Formula:** ETS = (Normalized_OS_Benefit) - (Normalized_AE_Increase)
- **Normalization Formula:** To make metrics comparable, normalize them on a scale from 0 to 1 based on the range observed across all included *experimental arms*.
  - Normalized_OS_Benefit = (OS_Arm - OS_Control) / (Max_OS_Benefit - Min_OS_Benefit)
  - Normalized_AE_Increase = (AE_Arm - AE_Control) / (Max_AE_Increase - Min_AE_Increase)
  - Where Max/Min_OS_Benefit and Max/Min_AE_Increase are the maximum and minimum differences observed between any experimental arm and its respective control *across all studies in the analysis*.

- **Sample Calculation:** Provide a full, step-by-step calculation for the ETS of the "100K Triplicate (Triplet Arm A)". Show the intermediate values for OS_Benefit, AE_Increase, the Max/Min range values, Normalized_OS_Benefit, Normalized_AE_Increase, and the final ETS.

Table 3: Pooled Clinical Metrics and Head-to-Head Scoring

R	C1: Study ID	C2: Study Type	C3: Trial Arm	C4: N	C5: Median OS (mo)	C6: OS vs Control (Δ mo)	C7: Grade ≥3 AEs (%)	C8: AEs vs Control (Δ %)	C9: Source URL	C10: Calculated ETS
R1	100K-Sim	Virtual	Triplet (Arm A)	20000	8.7	+2.6	94.0	+17.5	Report	[Calculated Value]
R2	100K-Sim	Virtual	Control (Arm E)	20000	6.1	0.0	76.5	0.0	Report	N/A
R3	100K-Sim	Virtual	Doublet (Arm D)	20000	[Calculated]	[Calc Δ]	N/R	N/A	Report	N/A
R4	MPACT	Real-World	Gem+Nab-P	[N]	8.5	[Calc Δ vs Gem]	84.0	[Calc Δ vs Gem]	[URL]	[Calculated Value]
R5	MPACT	Real-World	Gemcitabine	[N]	6.7	0.0	[Value]	0.0	[URL]	N/A
R6	[StudyID]	[Type]	[Arm Name]	[N]	[Value]	[Calc Δ]	[Value]	[Calc Δ]	[URL]	[Calculated Value]
R7	...	...	...	...	...	...	...	...	...	...

Authors' conclusions

- **Summary of Findings:** Synthesize the results from Tables 1, 2, and 3. Discuss the clinical implications. How does the virtual trial's triplet regimen (Arm A) compare to real-world standards of care like FOLFIRINOX or Gem+Nab-P when considering both efficacy and toxicity (as quantified by the ETS)? Highlight the promise (or lack thereof) of the virtual doublet (Arm D). Discuss the methodological concordance (e.g., OS) and discordance (e.g., ECOG profile) and its impact on the translatability of the virtual findings.
- **Key Research Gaps and Future Directions:** Based on the analysis, identify critical gaps in in-silico cancer modeling. Use the table below to structure these findings.

Table 4: Identified Research Gaps and Recommendations

R	C1: Identified Gap / Limitation	C2: Evidence from Analysis	C3: Proposed Future Direction / Recommendation	C4: Potential Impact
R1	Patient Profile Realism	The 100K-Sim's ECOG profile was significantly healthier than RWD from MPACT/NAPOLI-1. (Source: Report, Table 2)	Incorporate real-world data distributions (e.g., from Flatiron, COTA) into the virtual patient generation process.	Improves the generalizability and predictive accuracy of simulation outcomes for real-world populations.
R2	Model Complexity and Dynamics	The exponential survival model in the 100K-Sim does not capture treatment discontinuation or dose modification. (Source: Report, Table 1)	Develop and validate more sophisticated models (e.g., agent-based models, QSP) that simulate patient journeys more mechanistically.	Enables prediction of not just if a patient responds, but how and why, and allows for testing adaptive trial designs.
R3	Biomarker Granularity & Implementation	The report notes a data discrepancy in KRAS definition (91% vs 5%), potentially mis-applying the drug effect. (Source: Report)	Future models must link specific drug effects to validated biomarkers with high precision and apply them only to the correct subgroup.	Increases the power of simulations to identify potent biomarker-drug combinations and inform patient selection strategies.
R4	Standardization of In-Silico Reporting	Comparator in-silico studies report heterogeneous metrics, making direct comparison difficult. (Source: Table 1)	Advocate for standardized reporting guidelines for in-silico trials, analogous to CONSORT for RCTs.	Enhances transparency, reproducibility, and the ability to perform robust meta-analyses like this one.

Appendices

- **Full Electronic Search String:** Provide the exact search string used for PubMed/other databases.
- **Data-Extraction CSV:** Provide a Markdown table formatted as a CSV file, containing the raw data used to generate Table 3. This ensures data is machine-readable for future visualizations.

Generated csv

StudyID,StudyType,Phase,TrialArm,N,Median_OS_mo,OS_HR_vs_SoC,Median_PFS_mo,PFS_HR_vs_SoC,Grade3_plus_AE_pct,URL
100K-Sim,Virtual,III-equiv,Triplet (Arm A),20000,8.7,∼0.69,N/R,N/R,94.0,Source: Report
100K-Sim,Virtual,III-equiv,Control (Arm E),20000,6.1,1.00,3.1,1.00,76.5,Source: Report
[...populate with all other arms and studies from the analysis...]

"Start Report" "End Report"

[S57.REP.01.P43]

Table 24: Ref: S58.REP.02.P44

Meta-Analysis Charts: Prompt 44b

Based on the meta-analysis comparing the 100K-patient virtual PDAC trial with other virtual and real-world clinical trials, generate 10 visualizations with white backgrounds using 10 separate Python scripts that effectively communicate the following key comparisons and findings:

Create the following visualizations to highlight the critical insights from this comparative analysis:

01) Forest Plot: Display hazard ratios with 95% confidence intervals for overall survival comparing all experimental arms (virtual Triplet, virtual Doublet, MPACT Gem+nab-P, NAPOLI-1, FOLFIRINOX) versus their respective controls, showing how the virtual trial outcomes align with real-world trials
02) Scatter Plot with Efficiency Frontier: Plot median overall survival (x-axis) versus Grade ≥3 adverse events percentage (y-axis) for all treatment arms, with point sizes proportional to sample size and an efficiency frontier curve showing optimal efficacy-toxicity balance
03) Grouped Bar Chart: Compare median overall survival months across all experimental arms grouped by study type (virtual vs real-world), with error bars and control arm baselines shown as horizontal reference lines
04) Waterfall Plot: Display the Efficacy-Toxicity Score (ETS) for each experimental regimen ranked from highest to lowest, with positive scores in green and negative in red to show which treatments offer favorable risk-benefit profiles
05) Stacked Bar Chart: Show the distribution of ECOG performance status (0, 1, 2) across different trials to highlight the patient population discrepancy between the virtual trial (>95% ECOG 0-1) and real-world trials
06) Heatmap: Create a comparison matrix showing key metrics (OS benefit, PFS benefit, HR, Grade ≥3 AE increase, ETS) across all experimental arms with color intensity indicating magnitude of effect
07) Butterfly Plot: Display OS benefit (months gained) on the right and toxicity increase (% Grade ≥3 AE increase) on the left for each experimental arm, creating a mirror effect to visualize the trade-offs
08) Radar Chart: Compare the virtual Triplet, virtual Doublet, and FOLFIRINOX across multiple dimensions (OS, PFS, toxicity, patient fitness, biomarker specificity) to show their relative strengths and weaknesses
09) Sankey Diagram: Illustrate patient flow from baseline characteristics through treatment arms to outcomes, showing how the KRAS G12C subgroup (Archetype-05) derives enhanced benefit from Daraxonrasib-containing regimens
10) Timeline Visualization: Create a horizontal timeline showing the evolution of PDAC treatment standards from 2010-2025, marking when each real trial was conducted and where the virtual trial fits in the therapeutic landscape with median OS values annotated.

"Start Meta-analysis" "End Meta-analysis"

[S58.REP.02.P44]

Table 25: Ref: S58b.VIS.01.P44b

Financial Assessment: Prompt 45 (I/IV)

PROMPT FOR FINANCIAL IMPLICATIONS ASSESSMENT

Primary Instruction: You are an AI model specializing in life sciences finance and bioinformatics. Your task is to generate a complete, investment-grade "Financial Assessment and Value Proposition of a 100,000-Patient Triplicate Virtual Trial for PDAC Drug Development." This report must be framed for a startup seeking grant funding.

The analysis will focus on the financial and strategic value of the specific triplicate simulation methodology detailed in the provided reports (S57.REP.01.P43) compared to both alternative in-silico approaches and traditional in-person clinical trials. The clinical context will be drawn from the provided meta-analysis (S58.REP.02.P44).

Your output must be a single, plain-text document suitable for Google Docs. Use large, interpretable markdown tables with the specified R1, R2... and C1, C2... format for all structured data. You must not draw any final conclusions. Your role is to present the data, calculations, and financial frameworks as instructed, allowing the reader (e.g., a grant committee) to draw their own conclusions.

I. Executive Summary (Structured, ≤350 words)

Instruction: Generate a structured executive summary with the following sections:

- **Purpose:** State the report's purpose is to financially assess a triplicate virtual trial methodology as a capital-efficient tool for de-risking PDAC drug development for a startup.
- **Methodology:** Briefly describe the comparison of the 100K patient triplicate trial's costs and projected value against industry benchmarks for single-run virtual trials and Phase II/III in-person trials, using metrics like Cost of Evidence, De-Risking Value, and potential ROI.
- **Key Financial Findings (Instructions):** Synthesize the core financial arguments. For example: "The triplicate simulation, costing approximately $36,330 (Source: S57.REP.01.P43 costs), generated robust, verifiable evidence in 30 days. This represents a >99% cost reduction and a 98% timeline acceleration compared to a typical Phase III PDAC trial, which can exceed $100M and 5 years." Mention the value of identifying the superior risk-profile of Arm D as a key financial insight.
- **Value Proposition for Funding:** Frame the core argument for grant funding. For example: "This methodology provides a low-cost, high-confidence platform for making go/no-go decisions, preserving capital and directing resources toward assets with the highest probability of success. The robust, verifiable nature of the triplicate run (Source: S57.REP.01.P43, Table 04) is a key differentiator that minimizes investment risk."

II. Background: The Economic Imperative for Innovation in Oncology Trials

Instruction: Briefly describe the unsustainable economics of traditional oncology clinical trials, focusing on PDAC. Highlight the high cost, long duration, and >90% failure rate of drugs entering clinical phases. Frame in-silico trials not just as a scientific tool, but as a crucial financial strategy for startups to maximize capital efficiency and attract investment by generating early, robust evidence.

III. Objectives

Instruction: State the primary objectives of the financial assessment:

1. To quantify the direct and estimated costs of the 100K patient triplicate simulation.
2. To analyze the specific financial value and justification for the triplicate methodology versus a single simulation run.
3. To compare the "Cost of Evidence" from this virtual trial against estimated costs for other in-silico and traditional in-person PDAC trials.
4. To model the potential Return on Investment (ROI) and Net Present Value (NPV) of using this methodology to de-risk a drug development program, providing a quantitative basis for a grant application.

IV. Methods for Financial Assessment

Instruction: Detail the financial methodology.

- **Data Sources:** Primary financial data for the 100K patient triplicate simulation is extracted from S57.REP.01.P43. Clinical context and real-world trial outcomes are from S58.REP.02.P44. External financial benchmarks for comparator trials will be sourced from credible, citable industry reports and publications (e.g., from Tufts CSDD, BIO, Deloitte).
- **Financial Metrics:** List the key financial metrics that will be calculated and compared:
    - Total Project Cost (broken down into labor, compute, and third-party services)
    - Cost per Virtual Patient
    - Cost of Reproducibility (the marginal cost of the 2nd and 3rd runs)
    - Cost of Evidence (Total Cost / Key Actionable Insight)
    - Estimated Cost of Failure Avoidance
    - Burn Rate Reduction (Salaries and operational costs saved due to accelerated timeline)
    - Return on Investment (ROI)
    - Net Present Value (NPV) of Accelerated Development
- **Estimation Strategy:** State that when direct financial data for comparator trials is unavailable, estimates will be derived using established industry benchmarks. All assumptions, formulas, and sources for these estimates must be explicitly stated and justified. For example, labor costs for comparator virtual trials will be estimated based on reported team size, duration, and blended market-rate salaries for bioinformatics personnel.

V. Results

A. Cost-Benefit Analysis: Triplicate Simulation vs. Single-Run Virtual Trials

Instruction: Present a detailed cost breakdown of the 100K patient triplicate trial and compare it to estimated costs for other hypothetical single-run virtual trials. The purpose is to highlight the startup's operational efficiency and justify the cost of the triplicate methodology.

Table 26: Ref: S59.REP.03.P45

Table 1: Financial & Methodological Comparison of In-Silico Trial Methodologies
C1: Metric
C2: 100K Patient Triplicate Simulation
C3: Estimated Single-Run Virtual Trial (Standard)
C4: Estimated Advanced Mechanistic Model (e.g., QSP)
R1
Total Project Cost (USD)
(Calculate from S57.REP.01.P43 data)
(Estimate based on industry averages)
(Estimate based on higher complexity)
R2
Researcher Labor Cost
(Calculate from S57.REP.01.P43 data)
(Estimate: e.g., 2 researchers x 3 months x $120/hr)
(Estimate: e.g., 4 researchers x 6 months x $150/hr)
R3
AI/Cloud Compute Cost
(Sum from S57.REP.01.P43 data: $340)
(Estimate: e.g., $1,000 - $5,000)
(Estimate: e.g., $20,000 - $100,000+)
R4
Total Project Duration
30 days (Source: S58.REP.02.P44, Abstract)
(Estimate: 3-6 months)
(Estimate: 6-12 months)
R5
Cost of Reproducibility
(Calculate marginal cost of runs 2 & 3, likely dominated by compute/API costs)
Not Applicable (single run)
Not Applicable (single run)
R6
Cost per Virtual Patient
(Calculate: Total Cost / 100,000 patients)
(Calculate: Estimated Cost / Typical N, e.g., 1,000)
(Calculate: Estimated Cost / Typical N, e.g., 100)
R7
Key Methodological Benefit
High-confidence, verifiable results via triplicate runs and multi-AI validation (Source: S57.REP.01.P43, Table 04)
Rapid hypothesis screening
Deep biological mechanism exploration
R8
Source of Data/Estimate
S57.REP.01.P43, S58.REP.02.P44
(Cite industry report URL for labor/compute estimates)
(Cite industry report or publication URL for QSP cost estimates)

Sample Calculations for Section A:
Instruction: Provide three fully-worked sample calculations to demonstrate the required methodology.
Total Project Cost for 100K Triplicate: Labor Cost (1 researcher * 60 hr/wk * 4 wk * $150/hr) + AI/API Costs (

1. $260 + $30 + $20 + $20. Show the full calculation.
2. **Cost of Reproducibility:** Assuming labor was for the entire project, the marginal cost of the 2nd and 3rd runs is primarily the compute/API cost. Estimate this by assuming the initial run cost 1/3 of the total API cost, so the cost of reproducibility is (2/3) * $330. Justify this assumption.
3. **Estimated Labor Cost for Comparator (C4):** Based on a cited report that advanced QSP models require a team of 4 FTEs for 6 months, calculate the labor cost. (4 researchers * 24 weeks * 40 hr/wk * $150/hr). State all assumptions clearly.

Rationale for Estimates:
Instruction: Provide a paragraph justifying all assumptions made in Table 1 for columns C3 and C4. Cite sources for market-rate salaries, typical team sizes, and cloud computing costs for different types of in-silico projects.
B. Value Proposition: Capital Efficiency vs. Traditional In-Person Trials//
Instruction: Frame the financial comparison against traditional trials as a clear value proposition for a startup. Focus on capital preservation and risk reduction.

Table 27: Ref: S59.REP.03.P45

Table 2: Capital Efficiency and De-Risking: Virtual Triplicate vs. In-Person PDAC Trials

C1: Financial Metric
C2: 100K Patient Triplicate Simulation
C3: Typical Phase II PDAC Trial (Estimate)
C4: Typical Phase III PDAC Trial (Estimate)
R1
Total Estimated Budget (USD)
(Value from Table 1)
(Estimate, e.g., $15M - $25M)
(Estimate, e.g., $80M - $150M)
R2
Total Project Duration
30 days
(Estimate, e.g., 2 - 3 years)
(Estimate, e.g., 4 - 6 years)
R3
Cost per Patient (USD)
(Value from Table 1)
(Calculate: Budget / N, e.g., $20M / 150 patients)
(Calculate: Budget / N, e.g., $100M / 800 patients)
R4
Capital at Risk (for go/no-go decision)
(Total budget from R1)
(Full budget from R1)
(Full budget from R1)
R5
Time-to-Decision Value
Generates go/no-go evidence in 1 month, saving years of burn rate.
Requires years of investment before a clear signal emerges.
Requires the largest and longest investment for a definitive result.
R6
Key Actionable Insight
Identified superior risk-profile of Arm D; confirmed high toxicity of Arm A (Source: S58.REP.02.P44, Conclusions).
Typically tests one hypothesis (e.g., one new drug vs SoC).
Confirms efficacy/safety for registration, but at maximum cost.
R7
Source of Estimate
S57.REP.01.P43
(Cite source for Phase II costs, e.g., BIO/Tufts CSDD report URL)
(Cite source for Phase III costs, e.g., JAMA/DiMasi et al. URL)

Sample Calculations for Section B:

Instruction: Provide three fully-worked sample calculations.

1. **Cost of Failure Avoidance:** A key insight from the simulation was the extreme toxicity (94% Grade ≥3 AEs) of the triplet (Arm A) (Source: S58.REP.02.P44, Table 2). Estimate the value of this finding by calculating the cost of a failed Phase II trial ($20M) minus the cost of the simulation. This represents capital saved.
2. **Burn Rate Reduction:** Assume a startup's monthly burn rate for a clinical team (e.g., 5 personnel + overhead) is $100,000. Calculate the total savings from getting a decision signal in 1 month versus waiting 2 years for a Phase II trial to read out. (24 months * $100,000/month) - Simulation Cost.
3. **Cost per Patient Comparison:** Directly compare the "Cost per Virtual Patient" from Table 1, C2, R6 with the estimated "Cost per Real Patient" for a Phase III trial from Table 2, C4, R3. Express the difference as a percentage reduction.

Rationale for Estimates:

Instruction: Provide a detailed paragraph justifying the estimated budgets, durations, and patient numbers for Phase II and III PDAC trials in Table 2. Cite multiple authoritative sources (e.g., reports from Tufts Center for the Study of Drug Development, BIO, or academic publications on trial costs) to triangulate a credible range.

C. Investment Thesis: ROI and Grant Funding Justification

Instruction: Synthesize the previous analyses into a compelling investment thesis. Focus on how this specific triplicate methodology creates quantifiable value and serves as a prudent use of grant funds.

Table 28: Ref: S59.REP.03.P45

Table 3: Grant Funding Justification Framework

C1: Value Driver & Justification
C2: Key Supporting Finding from Simulation
C3: Quantifiable Financial Impact / Startup Value
C4: Source of Finding
R1
Optimizing Clinical Trial Design
(Value of designing a better, more successful trial)
The simulation confirmed a strong benefit for the KRAS G12C subgroup (Archetype-05).
This justifies a biomarker-driven trial design, which increases the probability of success (PoS). An increase in PoS from 10% to 30% on a $20M trial has a risk-adjusted value.
S57.REP.01.P43, Key Insights
R2
Justifying the Triplicate Methodology
(Value of robust, defensible evidence)
Cross-trial consistency scores were exceptionally high (avg. >8.5/10), and the multi-AI verification confirmed result stability.
This provides auditable, investment-grade evidence that reduces grantor risk. The marginal cost of the triplicate run is negligible compared to the increased confidence in the go/no-go decision.
S57.REP.01.P43, Table 04
R3
Accelerating Time-to-Market
(Value of speed)
The entire project was completed in 30 days, versus the 3-5 years required for an equivalent real-world evidence base.
An accelerated timeline brings potential revenue forward. The Net Present Value (NPV) of future cash flows increases significantly if they are realized 3 years earlier.
S58.REP.02.P44, Abstract
R4
Informing Future R&D
(Value of learning from model limitations)
The model's ECOG profile mismatch was identified as a key failure in external validation.
This is a critical, low-cost insight that informs the next, more accurate iteration of the simulation platform, improving its predictive power and future value to the startup's pipeline.
S57.REP.01.P43, Table 04

Sample Calculations for Section C:

Instruction: Provide three distinct, investment-focused sample calculations.

1. **Basic ROI of De-Risking:** Calculate the ROI based on the Cost of Failure Avoidance. ROI = [(Cost of Failed Phase II Trial - Cost of Simulation) / Cost of Simulation] * 100%. Use figures from previous sections.
2. **Net Present Value (NPV) of Acceleration:** Assume a potential drug has peak sales of $500M, 10 years from now. Show the NPV calculation using a discount rate (e.g., 15%). Then, re-calculate the NPV assuming the timeline is accelerated by 2 years (i.e., sales start in year 3 instead of year 5). The difference in NPV is the value of acceleration. Provide the formula: NPV = Σ [Cash Flow / (1 + r)^t].
3. **Valuation Uplift from Increased PoS:** A startup's pre-clinical asset might be valued at $5M. Industry data suggests that a successful Phase I/II result can increase valuation to $50M. If the simulation data increases the Probability of Success (PoS) for the Phase II trial from a baseline 10% to 25%, calculate the increase in the risk-adjusted asset value. Formula: Δ Value = (New PoS - Old PoS) * (Post-Phase II Valuation - Investment Cost).

Rationale for Estimates:

Instruction: Justify all assumptions used in the ROI and NPV calculations. Specifically explain the choice of discount rate, the estimated cost of a failed trial, and the basis for the Probability of Success figures, citing relevant financial or industry sources.

VI. Appendices

A. Data Extraction for Financial Modeling

Instruction: Create a CSV-formatted table that a financial analyst could use. Populate the first row with the 100K triplicate trial data. Leave subsequent rows as examples for comparator studies.

Data_Extraction_CSV

StudyID,StudyType,TotalBudget_USD_Est,Cost_per_Patient_USD_Est,Duration_Months,FTE_Count_Est,Primary_Financial_Value,Source_URL
PDAC-SIM-001_Triplicate,Virtual,36330,0.36,1,1,"De-risking of Arm A vs Arm D",S57.REP.01.P43
Comparator_Virtual_01,Virtual,,,,,,
Comparator_PhaseII_01,In-Person,,,,,,
Comparator_PhaseIII_01,In-Person,,,,,,
"Start Meta-Analysis: S58.REP.02.P44" "End Meta-Analysis: S58.REP.02.P44"
"Start Triplicate Results: S57.REP.01.P43" "End Triplicate Results: S57.REP.01.P43"
[S58.REP.02.P44] meta-analysis
[S57.REP.01.P43] triplicate runs

Table 29: Ref: S59.REP.03.P45

Financial Assessment Charts: Prompt 45b

Based on the included Financial Assessment and Value Proposition document, create 10 data visualizations with white backgrounds using separate Python scripts. Each visualization should clearly communicate key financial metrics and comparisons from the assessment to support grant funding decisions.

Generate the following 10 visualizations:

01) Horizontal Bar Chart: Compare total project costs across 100K triplicate simulation ($36,330), single-run virtual trial (~$120,000), QSP model (~$600,000), Phase II trial (~$20M), and Phase III trial (~$100M) using logarithmic scale to show order-of-magnitude differences
02) Stacked Bar Chart: Show cost breakdown of the 100K triplicate simulation between labor costs (~$36,000) and AI/cloud compute costs (~$330) to highlight the minimal infrastructure requirements
03) Timeline Comparison Chart: Display project duration in months for all five methodologies (triplicate: 1 month, single-run: 3-6 months, QSP: 6-12 months, Phase II: 24-36 months, Phase III: 48-72 months) as a horizontal timeline
04) Cost Per Patient Comparison: Create a bubble chart showing cost per patient on log scale ($0.36 virtual vs $133,000 Phase II vs $125,000 Phase III) with bubble size representing total patient count
05) ROI Waterfall Chart: Illustrate the 55,000% ROI calculation showing initial investment ($36,330), avoided failure cost ($20M), and net benefit as sequential steps
06) NPV Impact Visualization: Show the $39M value gain from 2-year acceleration using discount curves at 15% rate comparing $500M at year 8 vs year 10
07) Probability of Success Impact: Display before/after PoS (10% to 25%) and corresponding asset valuation increase ($5M to $9.5M) as paired bar charts
08) Capital at Risk Comparison: Create a risk matrix plot showing capital at risk vs time-to-decision for each methodology with bubble size representing uncertainty level
09) Cost of Reproducibility Analysis: Show marginal cost breakdown for triplicate methodology with first run cost vs additional runs ($220 for runs 2&3) highlighting confidence gain per dollar spent
10) De-risking Value Dashboard: Create a 2x2 grid showing four key metrics - cost savings from avoiding Arm A failure ($19.96M), burn rate reduction ($2.36M), cost reduction percentage (99.9997%), and ROI percentage (55,000%) as large-font metric cards.

“Start Financial Assessment and Value Proposition” “End Financial Assessment and Value Proposition”

[S59.REP.03.P45]

Table 30: Ref: S59b.VIS.01.P45b

References

[1] Kevin Kawchak. End-to-end pancreatic ductal adenocarcinoma digital twin clinical trial proposals, June 2025. doi:10.5281/zenodo.15735068.

[2] Kevin Kawchak. 10 year glioblastoma clinical trial meta-analyses by autonomous ai at scale. survival, hr, ae, and rob scored in ai reports and charts, including verifications, May 2025. doi:10.5281/zenodo.15549831.

[3] Kevin Kawchak. Ai revolution toward the cure of lung adenocarcinoma, April 2025. doi:10.5281/zenodo.15278152.

[4] Kevin Kawchak. mab bioprocess engineering in-context table forecasts using conversational ai literature insight generations. *ChemRxiv*, December 2024. doi:10.26434/chemrxiv-2024-jzbj0.

[5] Kevin Kawchak. Monoclonal antibody bioprocess engineering advancements using conversational artificial intelligence. *ChemRxiv*, October 2024. doi:10.26434/chemrxiv-2024-3m7m1.

[6] Kevin Kawchak. Paclitaxel biosynthesis ai breakthrough. *ChemRxiv*, October 2024. doi:10.26434/chemrxiv-2024-pqjd3.

[7] susan. New phase iii clinical trial actively recruiting, April 2025. URL: https://letswinpc.org/research/new-phase-iii-clinical-trial-actively-recruiting/.

[8] Arcus Biosciences. Data from a phase 1b study of quemliclustat-based regimens showed promising overall survival in treatment-naïve metastatic pancreatic cancer, 2024. URL: https://tinyurl.com/njm6scfm.

[9] Arsen Osipov, Ognjen Nikolic, Arkadiusz Gertych, Sarah Parker, Andrew Hendifar, Pranav Singh, Darya Filippova, Grant Dagliyan, Cristina R. Ferrone, Lei Zheng, Jason H. Moore, Warren Tourtellotte, Jennifer E. Van Eyk, and Dan Theodorescu. The molecular twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients. *Nature Cancer*, 5(2):299–314, February 2024. URL: https://www.nature.com/articles/s43018-023-00697-7, doi:10.1038/s43018-023-00697-7.

[10] Uzma Saddia Asghar, Matthew Griffiths, Amanzhol Kubeyev, Jordan Laurie, Andrea Giorni, Luiz A Zillmann Da Silva, Prabu Sivasubramaniam, Matthew Foster, and Andrew V. Biankin. Prediction of therapeutic response and cancer outcomes in solid tumours via in silico clinical trials. *Journal of Clinical Oncology*, 42:e13603–e13603, June 2024. URL: https://ascopubs.org/doi/10.1200/JCO.2024.42.16_suppl.e13603, doi:10.1200/JCO.2024.42.16_suppl.e13603.

[11] Hiroyuki Sayama, Kota Toshimoto, Georgi Kapitanov, Soumya Rao, Fei Hua, Kalyanasundaram Subramanian, Karyn Sutton, and Masayo Oishi. Comprehensive human quantitative systems pharmacology models for immuno-oncology: Enhancing decision-making on target indications and combination strategies for new drug development. In *Proceedings of the American Conference of Pharmacometrics 2024*, page 129, Phoenix, Arizona, USA, November 2024. International Society of Pharmacometrics. URL: https://tinyurl.com/33p73pnn, doi:10.70534/SYGH7193.

[12] Francesca Sarno, Jair Tenorio, Sofia Perea, Laura Medina, Roberto Pazo-Cid, Ignacio Juez, Rocío Garcia-Carbonero, Jaime Feliu, Carmen Guillen-Ponce, Pedro P. Lopez-Casas, Carmen Guerra, Yolanda Duran, José Francisco López-Acosta, Carolina Alonso, Estrella Esquivel, Ana Dopazo, Dipikaa Akshinthala, Senthil K. Muthuswamy, Pablo Lapunzina, Bruno Bockorny, and Manuel Hidalgo. A phase iii randomized trial of integrated genomics and avatar models for personalized treatment of pancreatic cancer: The avatar trial. *Clinical Cancer Research*, 31(2):278–287, January 2025. URL: https://aacrjournals.org/clincancerres/article/31/2/278/751097/A-Phase-III-Randomized-Trial-of-Integrated, doi:10.1158/1078-0432.CCR-23-4026.

[13] Andrew H. Ko, Do-Youn Oh, Janet Lau, Shivani K. Mhatre, Bo Ci, Robson Machado, Shi Li, Michael T. Bretscher, Irmarie Reyes-Rivera, Jiawen Zhu, Xiaosong Zhang, Jilpa Patel, Matthew A. Psioda, and Mariano Ponz-Sarvise. Investigational use of real-world data as a hybrid control in pancreatic ductal adenocarcinoma from the randomized phase ib/ ii morpheus trial. *Clinical Pharmacology Therapeutics*, 117(4):1021–1029, April 2025. URL: https://ascpt.onlinelibrary.wiley.com/doi/10.1002/cpt.3528, doi:10.1002/cpt.3528.

[14] Mohsen Pourmousa, Sankalp Jain, Elena Barnaeva, Wengong Jin, Joshua Hochuli, Zina Itkin, Travis Maxfield, Cleber Melo-Filho, Andrew Thieme, Kelli Wilson, Carleen Klumpp-Thomas, Sam Michael, Noel Southall, Tommi Jaakkola, Eugene N. Muratov, Regina Barzilay, Alexander Tropsha, Marc Ferrer, and Alexey V. Zakharov. Ai-driven discovery of synergistic drug combinations against pancreatic cancer. *Nature Communications*, 16(1):4020, April 2025. URL: https://www.nature.com/articles/s41467-025-56818-6, doi:10.1038/s41467-025-56818-6.

[15] Miksad RA Surinach A Corvino FA Wang S Torres AZ Mamlouk K Pulgar S Valderr Barzi, A. URL: https://tinyurl.com/mr45uw2v.

[16] OpenAI. Introducing openai o3 and o4-mini, 2025. URL: https://openai.com/index/introducing-o3-and-o4-mini/.

[17] OpenAI. Openai o3 and o4-mini system card. read the system card, 2025. URL: https://openai.com/index/o3-o4-mini-system-card/.

[18] Google AI Studio. We're expanding our gemini 2.5 family of models, june 2025. URL: https://blog.google/products/gemini/gemini-2-5-model-family-expands/.

[19] Google AI Studio. What will you build? push gemini to the limits of what ai can do, powered by the gemini api. this experimental model is for feedback and testing only. no production use., 2025. URL: https://aistudio.google.com/prompts/new_chat.

[20] xAI. Grok 4 is the most intelligent model in the world., 2025. URL: https://x.ai/news/grok-4.

[21] xAI. Grok 3 beta — the age of reasoning agents, 2025. URL: https://x.ai/news/grok-3.

[22] OpenAI. O3 is 80URL: https://community.openai.com/t/o3-is-80-cheaper-and-introducing-o3-pro/1284925.

[23] OpenAI. o3-pro | version of o3 with more compute for better responses, June 2025. URL: https://platform.openai.com/docs/models/o3-pro.

[24] Claude. Hybrid reasoning model that pushes the frontier for coding and ai agents, featuring a 200k context window. URL: https://www.anthropic.com/claude/opus.

[25] Claude. Introducing claude 4. URL: https://www.anthropic.com/news/claude-4.

[26] OpenAI. Openai flagship models. our versatile, high-intelligence flagship model. text and image input, text output. 128k context length. smarter model, higher price per token, 2024. URL: https://platform.openai.com/docs/models/gpt-4o.

[27] Claude. Claude sonnet 4 hybrid reasoning model with superior intelligence for high-volume use cases, and 200k context window. URL: https://www.anthropic.com/claude/sonnet.

[28] OpenAI. Introducing gpt-4.5, 2025. URL: https://openai.com/index/introducing-gpt-4-5/.

[29] Kevin Kawchak. Chatgpt 100,000 patient 24-month in silico phase iii 5-arm pancreatic cancer clinical trial triplicate, July 2025. doi:10.5281/zenodo.16415815.

12 Acknowledgments

The author would like to acknowledge OpenAI for providing access to ChatGPT, Google for providing access to Gemini, Anthropic for providing access to Claude, and xAI for providing access to Grok.

13 Ethical disclosures

The author of the article declares no competing interests.

14 Rights and permissions

15 About this study

Kawchak K. ChatGPT 100,000 Patient 24-Month In Silico Phase III 5-Arm Pancreatic Cancer Clinical Trial Triplicate. Zenodo. 2025; 10.5281/zenodo.16415815 [29].