
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

04.S35b.VER.03.P34

Prompt 34

The content is available under Creative Commons Attribution 4.0 International [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/).

ChatGPT o3-pro Research Generation: 12 Pages, July 10, 2025

Flatiron Validation Set:

<https://resources.flatiron.com/publications/real-world-outcomes-of-metastatic-pancreatic-cancer-mpc-patients-pts-treated-with-liposomal-irinotecan-nal-iri-in-the-us>

Real world outcomes of metastatic pancreatic cancer (mPC) patients (pts) treated with liposomal irinotecan (nal-IRI) in the US

“Trial 1”

Prompt 34: Log + Prompt

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 $\leq 5.0\%$, else "Fail". For R3, mark "Pass" if the C2 value is ≥ 0.950 , else "Fail".

3. Table T3 – ECOG Concordance (3 rows \times 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 $|C2 \text{ value} - C3 \text{ value}|$. For R2 (ECOG 1), this would be $|\text{Simulated \% for ECOG 1} - 60.0|$.

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 $\pm 5\%$

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]

Output:

Thanks for uploading the simulation log. Just to confirm before I process the data:

1. Should the OS estimates and ECOG calculations be restricted to any specific treatment arm (e.g., simulated control/Arm E), or should I include all arms in the analysis?
2. Is the Kaplan-Meier OS estimate based on raw proportions surviving at each monthly landmark, or do you want me to derive survival probabilities from a full KM curve fit?

Once I have this information, I'll proceed with generating all three tables and the interpretation summary.

Input:

Use provided prompt.

Generation 35b (Formatted)

ChatGPT o3-pro Research

483 words, 11 minutes 37.22 seconds

Got it. I'll analyze the simulation log using data from Arm E (simulated control) to generate the required validation tables and interpretation. I'll calculate Kaplan-Meier OS estimates at monthly landmarks, compare them with Flatiron reference values, evaluate ECOG distribution concordance, and provide validation status based on defined thresholds.

I'll let you know as soon as the validation report is ready.

Validation Report

Table T1 – OS Concordance

Month (mo)	Simulated OS %	Flatiron OS %	Absolute Difference %
0	100.0%	100.0%	0.0%
3	72.0%	70.0%	2.0%
6	51.3%	52.0%	0.7%
9	36.3%	40.0%	3.7%
12	25.7%	28.0%	2.3%
18	13.1%	15.0%	1.9%
24	6.7%	8.0%	1.3%

Example Calculation for Table T1: C4 (Absolute Difference %) is the absolute gap between simulated and Flatiron OS percentages at each time point. For example, at Month 3 (R2), the difference is $|72.0\% - 70.0\%| = 2.0\%$. All seven differences above (R1–R7) will be used to derive the standard deviation in Table T2.

Table T2 – OS Summary Metrics

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

Example Calculations for Table T2:

- R1 (Mean OS %): Calculate the average of simulated OS% from Table T1 months 3–24. For example, $(72.0\% + 51.3\% + 36.3\% + 25.7\% + 13.1\% + 6.7\%) / 6 = 34.2\%$.
- R2 (SD of differences): Using the seven absolute differences from Table T1 (0.0%, 2.0%, 0.7%, 3.7%, 2.3%, 1.9%, 1.3%), the sample standard deviation is $\approx 1.2\%$.
- R3 (Pearson r): Compute the correlation between the Sim OS% and Flatiron OS% vectors from Table T1. This yields $r \approx 0.999$ (to three decimals).

Table T3 – ECOG Concordance

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%

Example Calculation for Table T3: C4 (Absolute Difference %) is calculated as the absolute difference between simulated and reference ECOG proportions. For instance, for ECOG 1 (R2), $|73.9\% - 60.0\%| = 13.9\%$.

Interpretation

The simulated OS trajectory closely mirrors the Flatiron benchmarks: at each reported time point (0–24 mo), simulated survival falls within $\pm 5\%$ of the real-world OS (all points concordant). Summary metrics further confirm strong alignment (mean OS% difference $\sim 1.3\%$ – Pass; SD of differences $\sim 1.2\%$ – Pass; Pearson $r = 0.999$ – Pass). However, baseline ECOG distribution shows large discrepancies (simulated vs Flatiron differences of 5.2%, 13.9%, 19.1% for ECOG 0/1/2 respectively, each exceeding $\pm 5\%$ and failing concordance). **Overall**, the simulation’s survival outcomes demonstrate high external validity, but the mismatch in patient ECOG profile indicates a notable limitation in its representativeness.

Generation 35b (Plain Text)

ChatGPT o3-pro Research

483 words, 11 minutes 37.22 seconds

Got it. I'll analyze the simulation log using data from Arm E (simulated control) to generate the required validation tables and interpretation. I'll calculate Kaplan-Meier OS estimates at monthly landmarks, compare them with Flatiron reference values, evaluate ECOG distribution concordance, and provide validation status based on defined thresholds.

I'll let you know as soon as the validation report is ready.

Validation Report

Table T1 – OS Concordance

Month (mo)	Simulated OS %	Flatiron OS %	Absolute Difference %
0	100.0%	100.0%	0.0%
3	72.0%	70.0%	2.0%
6	51.3%	52.0%	0.7%
9	36.3%	40.0%	3.7%
12	25.7%	28.0%	2.3%
18	13.1%	15.0%	1.9%
24	6.7%	8.0%	1.3%

Example Calculation for Table T1: C4 (Absolute Difference %) is the absolute gap between simulated and Flatiron OS percentages at each time point. For example, at Month 3 (R2), the difference is $|72.0\% - 70.0\%| = 2.0\%$. All seven differences above (R1–R7) will be used to derive the standard deviation in Table T2.

Table T2 – OS Summary Metrics

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

Example Calculations for Table T2:

* R1 (Mean OS %): Calculate the average of simulated OS% from Table T1 months 3–24. For example, $(72.0\% + 51.3\% + 36.3\% + 25.7\% + 13.1\% + 6.7\%) / 6 = 34.2\%$.

* R2 (SD of differences): Using the seven absolute differences from Table T1 (0.0%, 2.0%, 0.7%, 3.7%, 2.3%, 1.9%, 1.3%), the sample standard deviation is $\approx 1.2\%$.

* R3 (Pearson r): Compute the correlation between the Sim OS% and Flatiron OS% vectors from Table T1. This yields $r \approx 0.999$ (to three decimals).

Table T3 – ECOG Concordance

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%

Example Calculation for Table T3: C4 (Absolute Difference %) is calculated as the absolute difference between simulated and reference ECOG proportions. For instance, for ECOG 1 (R2), $|73.9\% - 60.0\%| = 13.9\%$.

Interpretation

The simulated OS trajectory closely mirrors the Flatiron benchmarks: at each reported time point (0–24 mo), simulated survival falls within $\pm 5\%$ of the real-world OS (all points concordant). Summary metrics further confirm strong alignment (mean OS% difference $\sim 1.3\%$ – Pass; SD of differences $\sim 1.2\%$ – Pass; Pearson $r = 0.999$ – Pass). However, baseline ECOG distribution shows large discrepancies (simulated vs Flatiron differences of 5.2%, 13.9%, 19.1% for ECOG 0/1/2 respectively, each exceeding $\pm 5\%$ and failing concordance). **Overall,** the simulation’s survival outcomes demonstrate high external validity, but the mismatch in patient ECOG profile indicates a notable limitation in its representativeness.

Research Summary Generation 35b: 903 Words

Validation Report for Patient-Level Simulation Log

ChatGPT

I'm pulling together simulated and Flatiron OS percentages at specified months to construct a meaningful table. It involves calculating Kaplan-Meier survival estimates from the simulation log.

I'm combining Kaplan-Meier survival estimates with simulated control data, ensuring precise calculations for OS% to inform the analysis.

Searched for S33.TRL.13.P30.LOG.csv data

ChatGPT

I'm exploring the CSV file using Python, specifically accessing it via the local path "/mnt/data/S33.TRL.13.P30.LOG.csv", as Python has access to it.

Analyzed with python

ChatGPT

Piecing together patient ID, arm, archetype, and survival data. Os_event_flag isn't immediately clear. Exploring potential columns like death_flag or similar for precise mapping.

Analyzed with python

ChatGPT

Wondering if "time_to_death" means all will ultimately die, or if some might survive. Considering censoring and progression details for clear insights.

Analyzed with python

ChatGPT

OK, let me see: a simulation has 100,000 entries, with a median time-to-death of 7.35 months. This suggests better survival than the known median OS of 5.6 months.

Analyzed with python

ChatGPT

OK, let me break this down: The control group has 20k patients with a median death time of 6.21 months and a mean time of 8.867 months. Oxygen counts are 0, 1, and 2.

Analyzed with python

ChatGPT

I'm piecing together Kaplan-Meier survival curves to estimate OS% at key timepoints, considering factors like exact time events and potential immediate deaths.

Analyzed with python

ChatGPT

I'm piecing together survival rates at various time intervals. Considering patients passing at exact months, I'm mapping their counts to the total population.

I'm thinking about how survival function $S(t)$ includes events at time t , and whether or not to consider specific event times in Kaplan-Meier curves.

Analyzed with python

ChatGPT

Issued guidelines for classifying survival thresholds. Clarified the inclusion of deaths precisely at the threshold in survival probability calculations.

Analyzed with python

ChatGPT

Calculating the death rate at time 0—0.05% of 20,000 equals 10, suggesting 10 deaths at that moment.

Analyzed with python

ChatGPT

I'm comparing simulated survival rates to flatiron values at various time points. The differences are minor, all falling within a 5% threshold, which seems acceptable for this analysis.

Analyzed with python

ChatGPT

OK, let me see. Adding all exact simulation OS values to find their average, carefully rounding to one decimal point.

Analyzed with python

Analyzed with python

ChatGPT

I'm piecing together the sum of alive counts and checking the fraction using the given numbers. The calculated time-matched sum fraction looks good.

Analyzed with python

ChatGPT

I'm summing flatiron and sim differences and checking their standard deviation using sample principles, including differences at month 0. Progressing through differences at 0, 3, 6, 9, 12, 18, and 24 months.

Analyzed with python

ChatGPT

Calculating mean differences and standard deviations, ensuring accuracy and clarity. 결정하고 있는 시점에서 중심 차이와 표준편차를 계산하고 더 정확하고 명확한 분석을 하는 중입니다.

Analyzed with python

ChatGPT

I'm working through presenting OS data, checking rounding rules, and ensuring precision in percentages. Progress involves validating external validity with a clear summary of outcomes and discrepancies.

Assessing external validity reveals mixed results: survival outcomes align well while baseline ECOG distribution significantly diverges, suggesting partial reliability dependent on the aspect under consideration.

Analyzed with python

ChatGPT

OK, let me see. Ensuring tables align correctly, centering text, and right-aligning numeric columns for consistency. Each table setup is confirmed for clarity and accuracy.

Analyzed with python

ChatGPT

I'm thinking through Python's rounding method for percentages, specifically 20.15%. It rounds down to 20.1%, but half-up rounding might be more suitable, affecting accuracy.

Analyzed with python

ChatGPT

I'm mapping out the need for a decimal in Sim OS at 12 using 20.15 and comparing other sim values to pinpoint patterns. Decimal precision seems crucial. Progressing through these comparisons, I'm ensuring accuracy.

Analyzed with python

ChatGPT

I'm noting that the option has been confirmed.

Analyzed with python

ChatGPT

I'm ensuring all decimal values are accurate, with a focus on rounding to one decimal place for clarity and consistency in the validation report.

Analyzed with python

Analyzed with python

ChatGPT

I'm piecing together how using pandas series with Decimal turns out boolean results into a sum. It's interesting to see the conversion and percentage calculation process unfold, potentially producing decimals.

Analyzed with python

ChatGPT

I overlooked running the code after computing sim_os_vals, missing the block's finalization step. This likely means the code hasn't been fully executed.

Analyzed with python

ChatGPT

I'm mapping out if the environment handled print or flush issues, possibly due to DataFrame operations or incorrect type conversion. Decimal should manage both int and str efficiently.

Analyzed with python

ChatGPT

Hmm, I'm thinking about why printing isn't working. It could be due to output capture or a large list, which is odd since it's just seven lines.

Analyzed with python

ChatGPT

Is the issue due to buffering or a large, slow environment?

Analyzed with python

ChatGPT

Potential causes for the output not displaying include issues with the loop, the large timesE variable, or the Decimal in construct.

Analyzed with python

ChatGPT

Let's think about the data: it aligns with expectations.

Analyzed with python

ChatGPT

I'm mapping out the differences between percentages: 5.2%, 13.9%, and -19.1%.

I'm manually crafting the answer, avoiding AI-generated content or internet searches to ensure accuracy and relevance.

Research completed in 10m · 0 sources

Citations Generation 35b

All Sources