HopeAI

Evidence Based Clinical Trial Design

When the Power of AI Meets Statistical Rigor

Will Ma

Founder & CEO

When trials fail, design often fails first

IMMUNOLOGY

EFC11574 (sarilumab)

RA, post-Humira · Sanofi

$60M

spent before termination

DESIGN LESSON

Optimistic run-in assumption a simulation could have flagged

ONCOLOGY

CheckMate 459

1L HCC · Bristol Myers Squibb

p = 0.0752

vs threshold p ≤ 0.0419

DESIGN LESSON

Accidental covariate imbalance hid a real efficacy benefit

ONCOLOGY

Cylembio + pembro

1L melanoma · IO Biotech

p = 0.056

vs threshold p ≤ 0.045

DESIGN LESSON

Aggressive alpha boundary swallowed a real clinical signal

ONCOLOGY

LITESPARK-012

1L ccRCC · Merck / Eisai

Dual miss

PFS + OS at interim

DESIGN LESSON

Triplet on top of an already-active backbone — wrong line

ONCOLOGY

LATIFY

2L NSCLC · AstraZeneca

16.9 vs 4.4 mo

DOR in responders – yet no OS/PFS/ORR gain overall

DESIGN LESSON

Durable responder signal wash out across an unselected population

They failed at decision points where evidence was incomplete and design was suboptimal

THE PATTERN

57%

of phase 3 oncology trials report a hazard ratio weaker than the one assumed in their power calculation.

JNCI, 2025

ASSUMED VS REALITY

Inaccurate benchmarking

HIDDEN IN PLAIN SIGHT

Accidental covariate imbalances

WRONG POPULATION

All-comers, no enrichment

COMPOSITE OR SURROGATE

Endpoint mismatch

HopeAI – derisk clinical trials with evidence-based designs

Company

Founded in: **2023**

MAYO CLINIC PLATFORM

A Mayo Clinic Platform Accelerate company

Team Background

Bristol Myers Squibb

FDA

MERCK

Imperial College London

COLUMBIA UNIVERSITY

Valued Clients

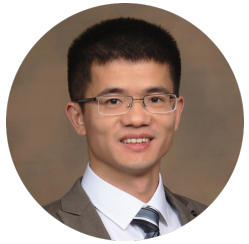
Debiopharm

Bristol Myers Squibb

Genentech

...

The team – deep pharma experience, AI-native engineering



Will Ma
CEO, Ex-Sanofi, Moffitt, BMS

Feifang Hu
CSO, adaptive design pioneer

Ram Tiwari
Regulatory, Ex-FDA Director

Tanja Obradovic
Chief Medical Strategy Officer, ex-Merck, Takeda

En Xie
CTO, Imperial College London

Zixuan Zhao
Head of Stat Innovation

Zexin Ren
AI Researcher

Xiaomai Zhang
Chief Marketing Officer

Bristol Myers Squibb

MERCK

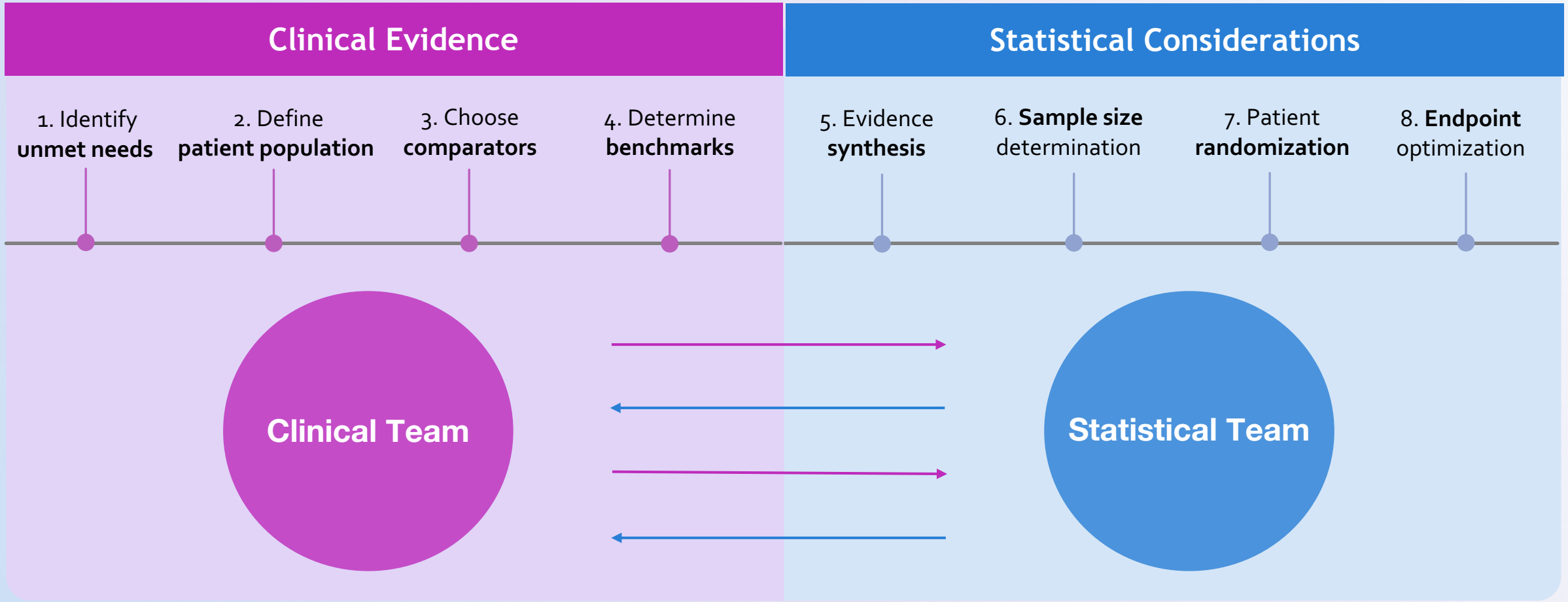
FDA

Imperial College London

GW

Columbia University

Clinical trial design: A lengthy, iterative process ready for AI acceleration



6-12 months of back-and-forth discussions, **$1M** in potential revenue delay **each day**

Four solutions. One trial-design AI engine.

PURE Evidence

SYSTEMATIC LITERATURE REVIEW

AI-enabled, expert-curated evidence. SLRs in days, not months.

SynthIPD

SYNTHETIC INDIVIDUAL PATIENT DATA

Reconstruct IPD from published KM curves — privacy-preserving, pixel-level accuracy.

AI Clinician

PROTOCOL & COHORT DESIGN

Evidence-based recommendations on endpoints, eligibility, and cohorts.

AI Statistician

TRIAL SIMULATION & ADAPTIVE DESIGN

Power, sample size, adaptive boundaries — across thousands of scenarios.

Outline

PURE Evidence – A semi-automated AI system for clinical evidence synthesis

AI Statistician – Innovative trial designs and no-code trial simulations

SynthIPD – Reconstructing patient-level data from published trials

Better evidence drives better designs
Better designs save trials

Motivating examples of evidence in trial design

Establish Control Benchmarks

Establish control benchmarks from prior trials to inform sample size estimation.

Association Analyses between multiple endpoints

Aggregate efficacy data across studies for association analyses between different endpoints.

Performance in specific subgroup

Establish aggregated treatment effect estimate within a specific subgroup.

PURE Evidence: SLRs in 1 week, not 6 months

Systematic literature review is the foundation of every study design — and the often-compromised step.

TRADITIONAL SLR

6 months

- ✕ Manual screening of thousands of papers
- ✕ Inconsistent extraction across reviewers
- ✕ Outdated by the time it's complete
- ✕ $12M+ per year per big pharma

→

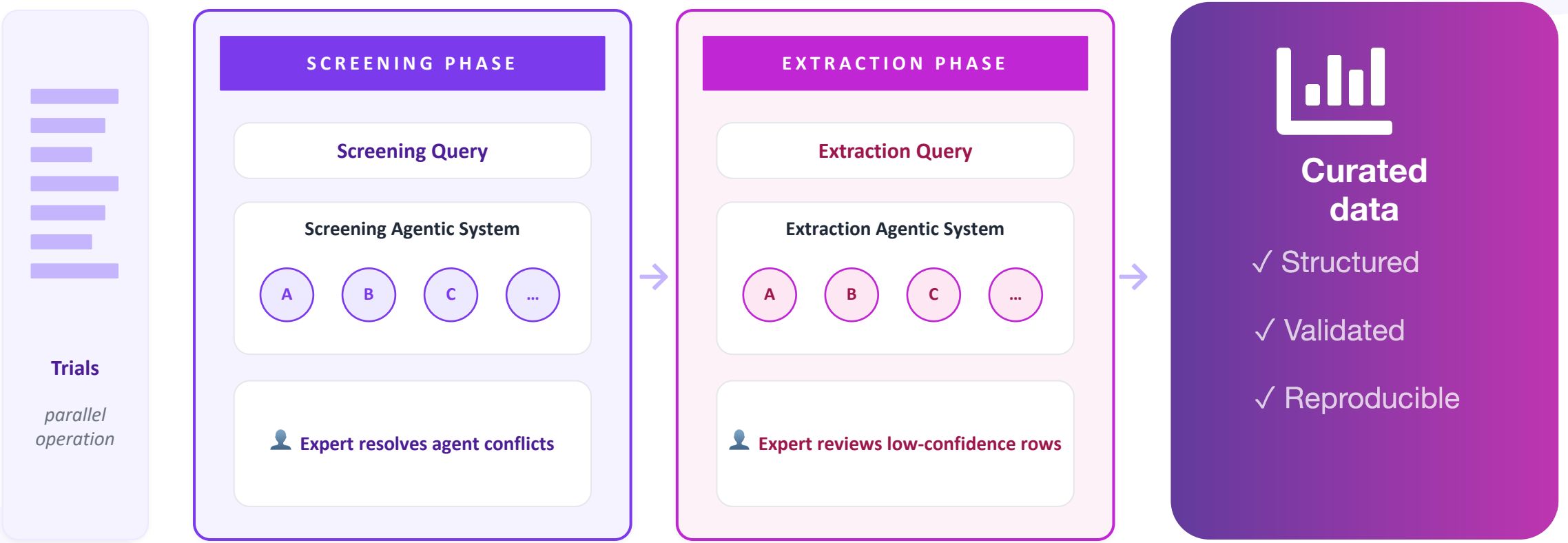
PURE EVIDENCE

1 week

- ✓ Agentic retrieval across PubMed + ClinicalTrials.gov
- ✓ Standardized study- and subgroup-level outcomes
- ✓ Continuously refreshed, audit-ready provenance
- ✓ Senior biostatistician QC on every deliverable

~20× faster than traditional SLR — at a fraction of the cost.

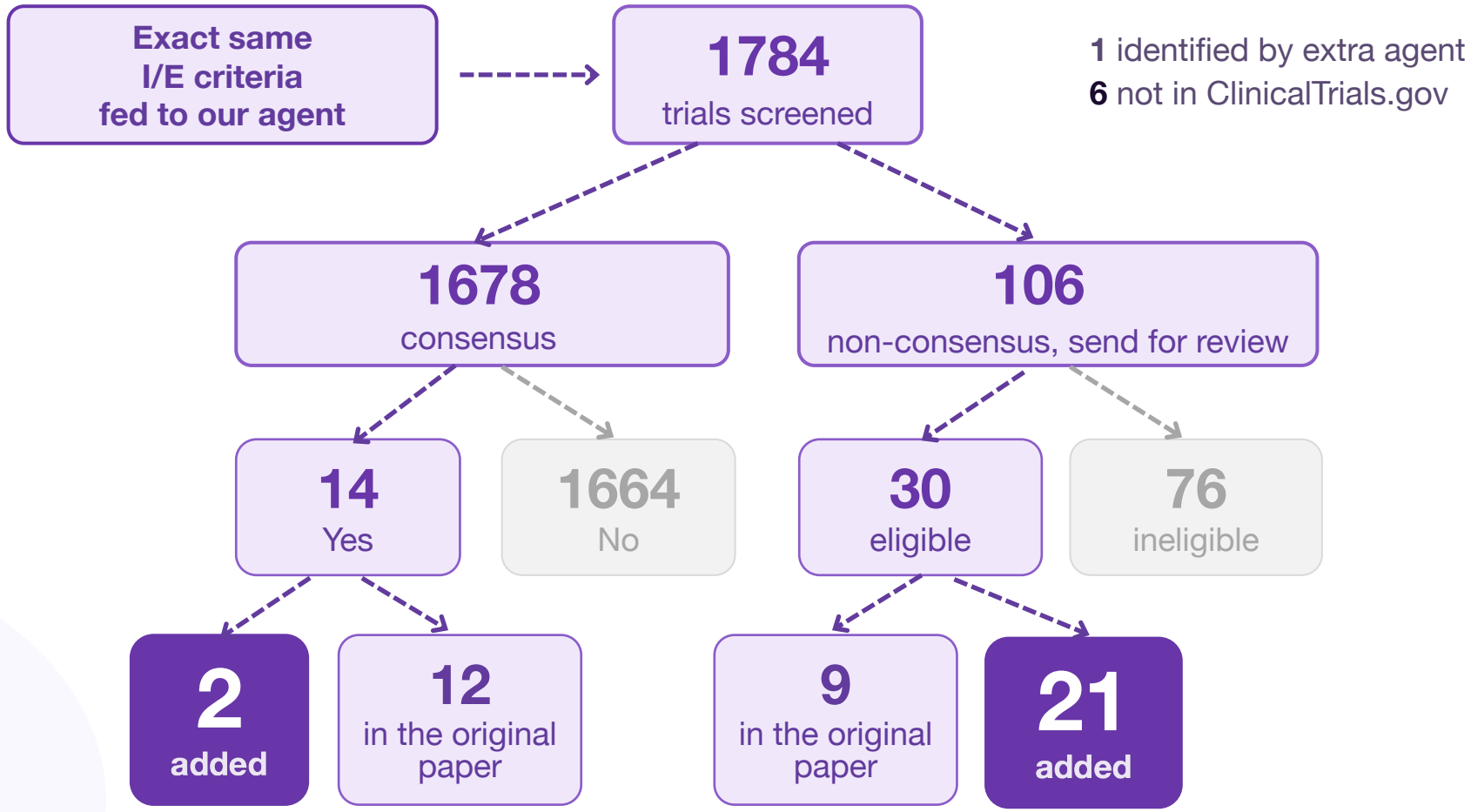
Two Multi-Agent Systems with Human-in-the-Loop



Critical feature: the explicit space for expert-in-the-loop at both phases.

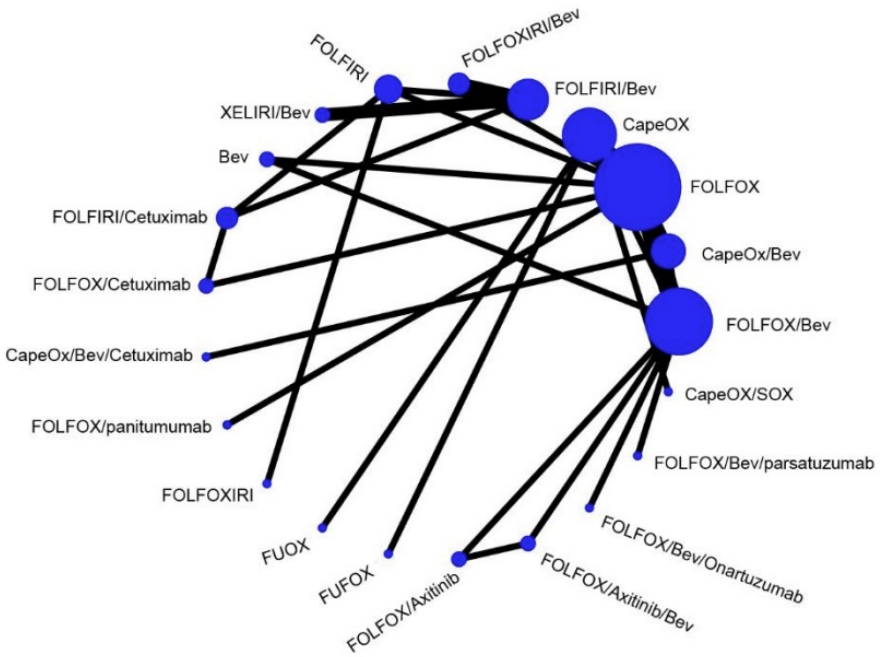
Reproducing a Published Network Meta-Analysis

Shan Xu, Ali Sak, and Yasin Bahadir Erol. Network meta-analysis of first-line systemic treatment for patients with metastatic colorectal cancer. Cancer Control, 28:10732748211033497, 2021.



Reproducing a Published Network Meta-Analysis

sFigure2: *Network diagram for each outcome*



2.A Overall survival. From the network geometries for OS, FOLFOX had the largest samples investigated in most arms compared to the control group, and thus its node edge was the largest, followed by FOLFOX/Bev and CapeOX. Each treatment held at least one direct evidence compared with the control group.

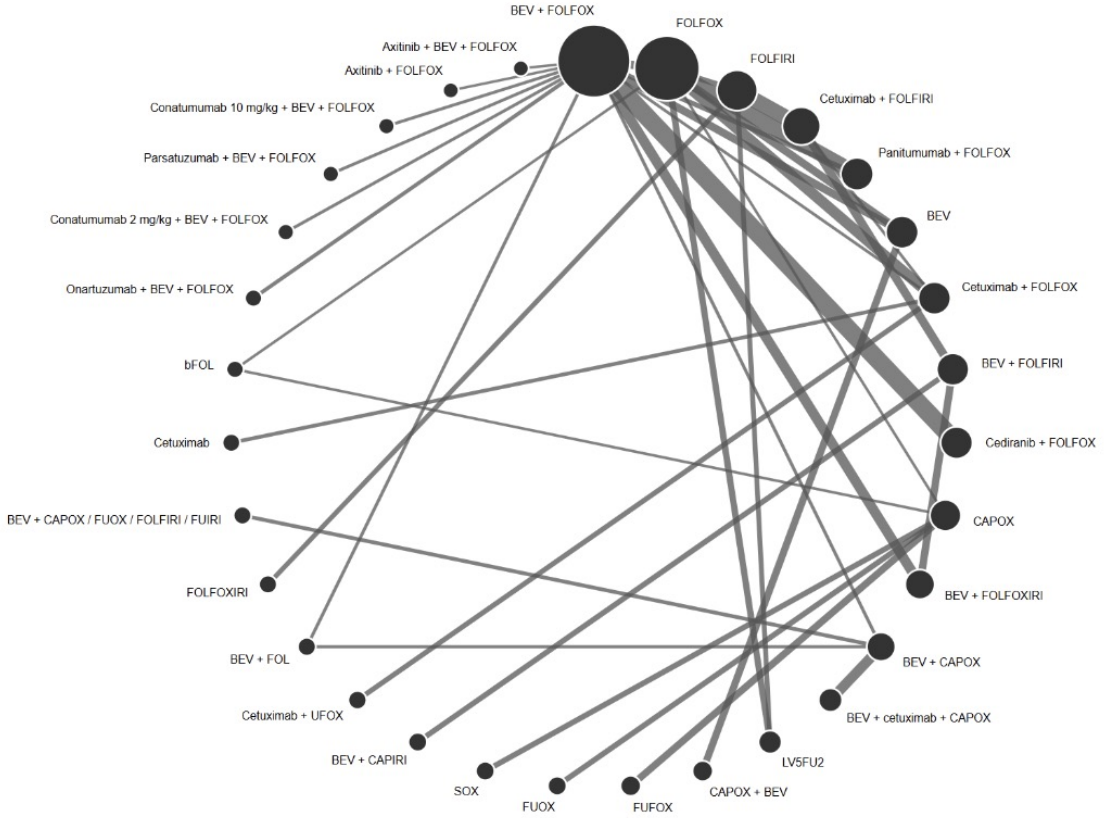
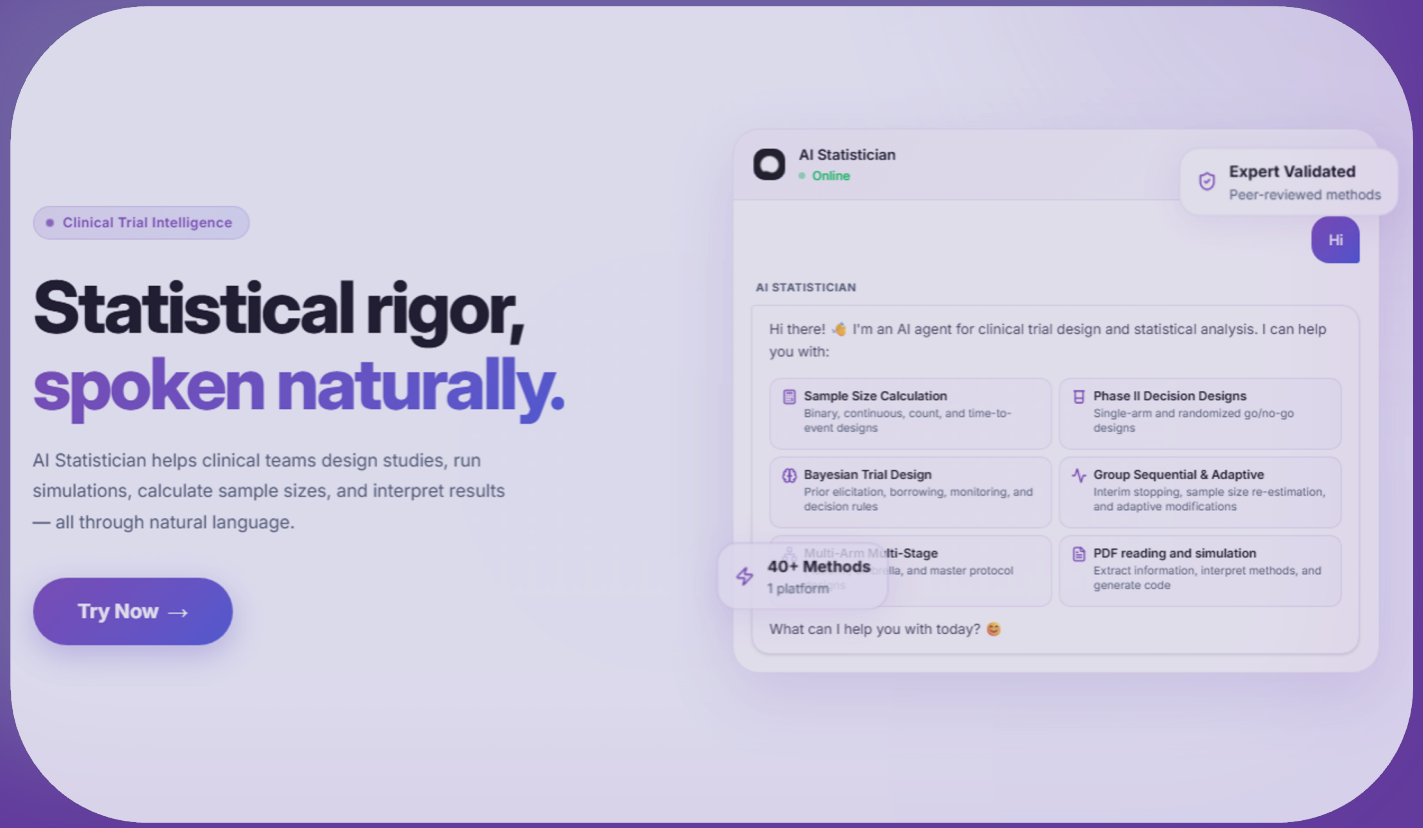


Figure 8: Network geometry plot for the included trials. Solid circles are proportional to total sample size. Line widths are proportional to each trial's (each comparison's) sample size. For the explanation of durg acronyms, see Tables 6, 7.

AI Statistician

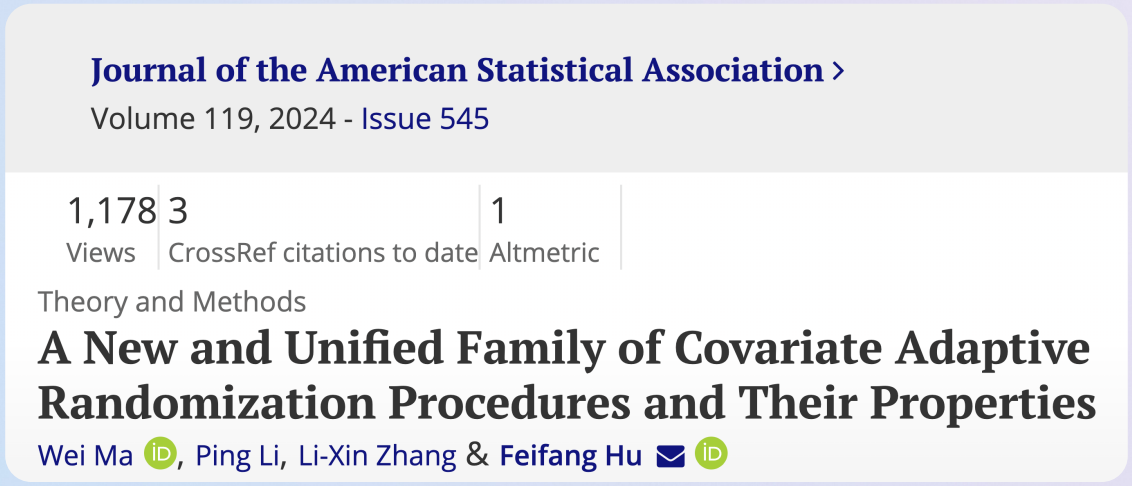
Your AI teammate for clinical trial design

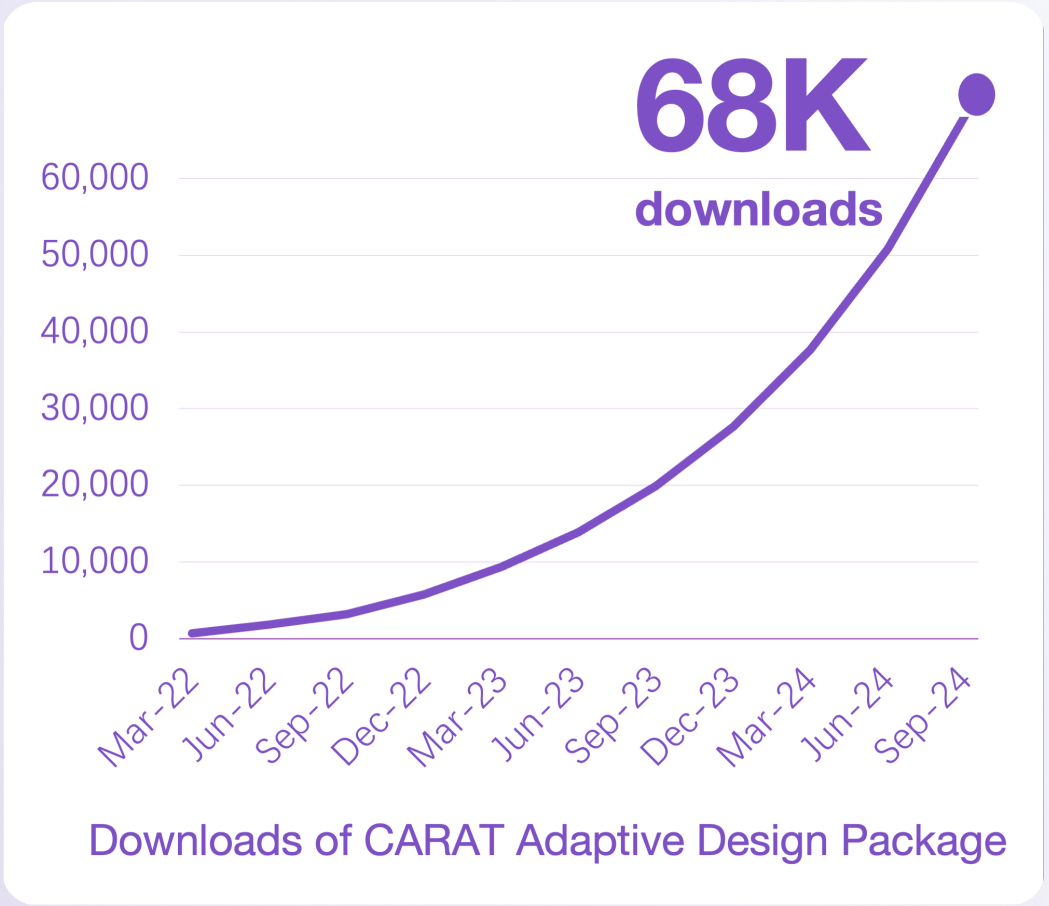


Leveraging decades of statistical innovations in trial design and data analysis

100+
research papers

2
FDA whitepapers

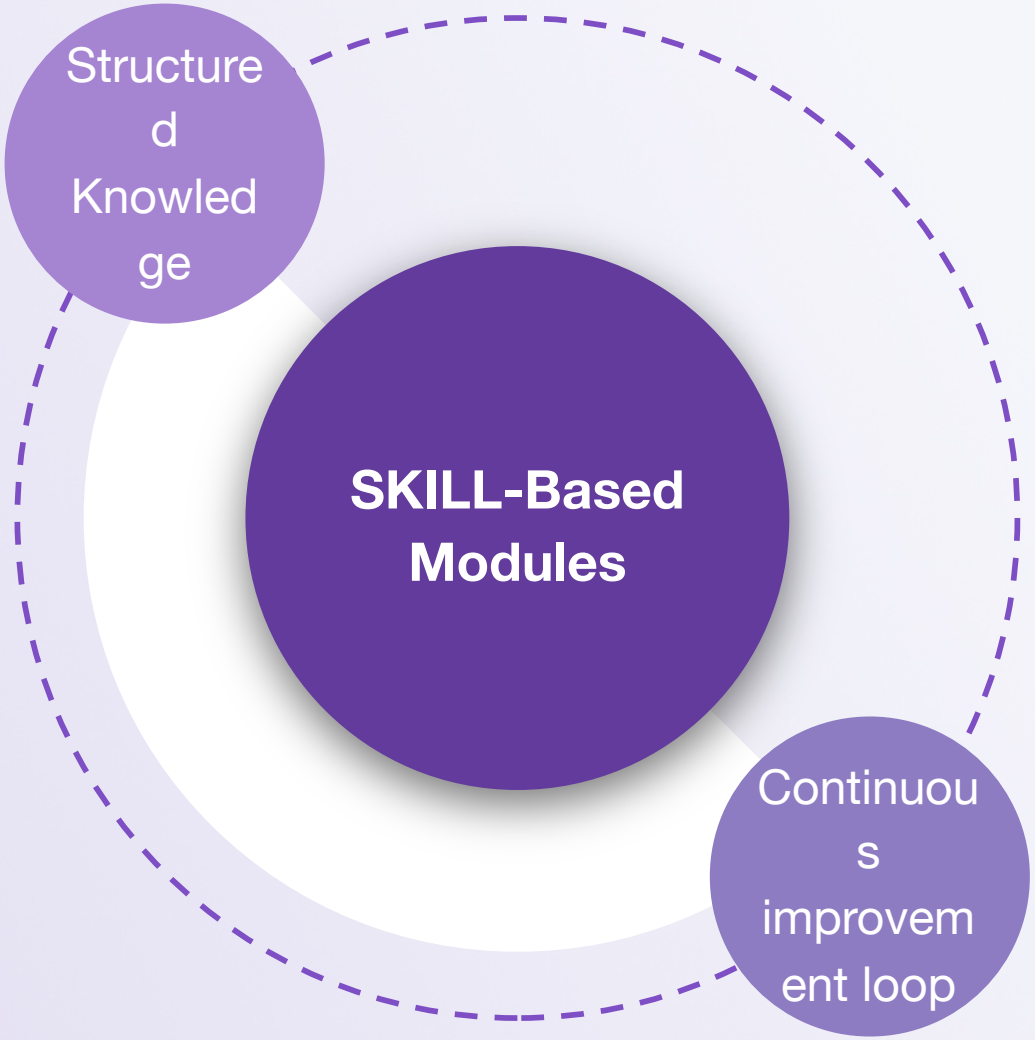




Skill Architecture

Each SKILL is a self-contained workflow module that packages set of instructions, code, references, and workflow assets that teaches the AI how to perform a specific task reliably.

```
skill-name/
  ├── SKILL.md        Main instructions and workflow logic
  ├── scripts/        Statistical code and simulation
programs
  ├── references/     Papers, R package methods
  ├── examples/       Example use cases
  └── experience/     Prior run results and validation
```



Equipping AI Statistician with SKILLs

Each SKILL encapsulates domain expertise into a modular, reusable unit. The AI Statistician composes multiple SKILLs to handle complex trial design workflows end-to-end.

How SKILLs Work

1 Interpret

The AI reads a user's question and selects the appropriate SKILL based on the task type.

2 Execute

The SKILL provides workflow instructions, validated code, and reference papers. The AI follows these steps to produce results.

3 Validate

Outputs are checked against built-in examples and prior results stored in the experience folder.

4 Learn

Successful runs are saved back into the SKILL, improving accuracy and reliability over time.

Building Statistician Capabilities

Evidence Extraction SKILLs

Query clinical databases, synthesize efficacy data, and extract design-relevant parameters from published studies.

Sample Size Calculator SKILLs

Compute sample sizes for frequentist and Bayesian designs with validated R scripts and simulation programs.

Adaptive Design SKILLs

Configure interim analyses, futility boundaries, and adaptive enrichment strategies using peer-reviewed methods.

Simulation SKILLs

Run trial simulations to evaluate operating characteristics, type I error control, and power across design scenarios.

Paper-to-Code SKILLs

Translate published statistical methods into executable, reproducible R or Python code for direct application.

Will AI Replace Statisticians?

A skill-by-skill classification for biostatisticians in biopharma

LIKELY REPLACED BY AI

- **Statistical programming**
  SAS / R / Python code generation & validation
- **Evidence synthesis**
  Literature extraction, efficacy data aggregation
- **Sample size calculation**
  Frequentist, Bayesian, simulation-based
- **SAP drafting**
  Standard sections, templates, formatting
- **Data visualization**
  Table shells, figure recommendations

AI AUGMENTS — HUMAN DECIDES

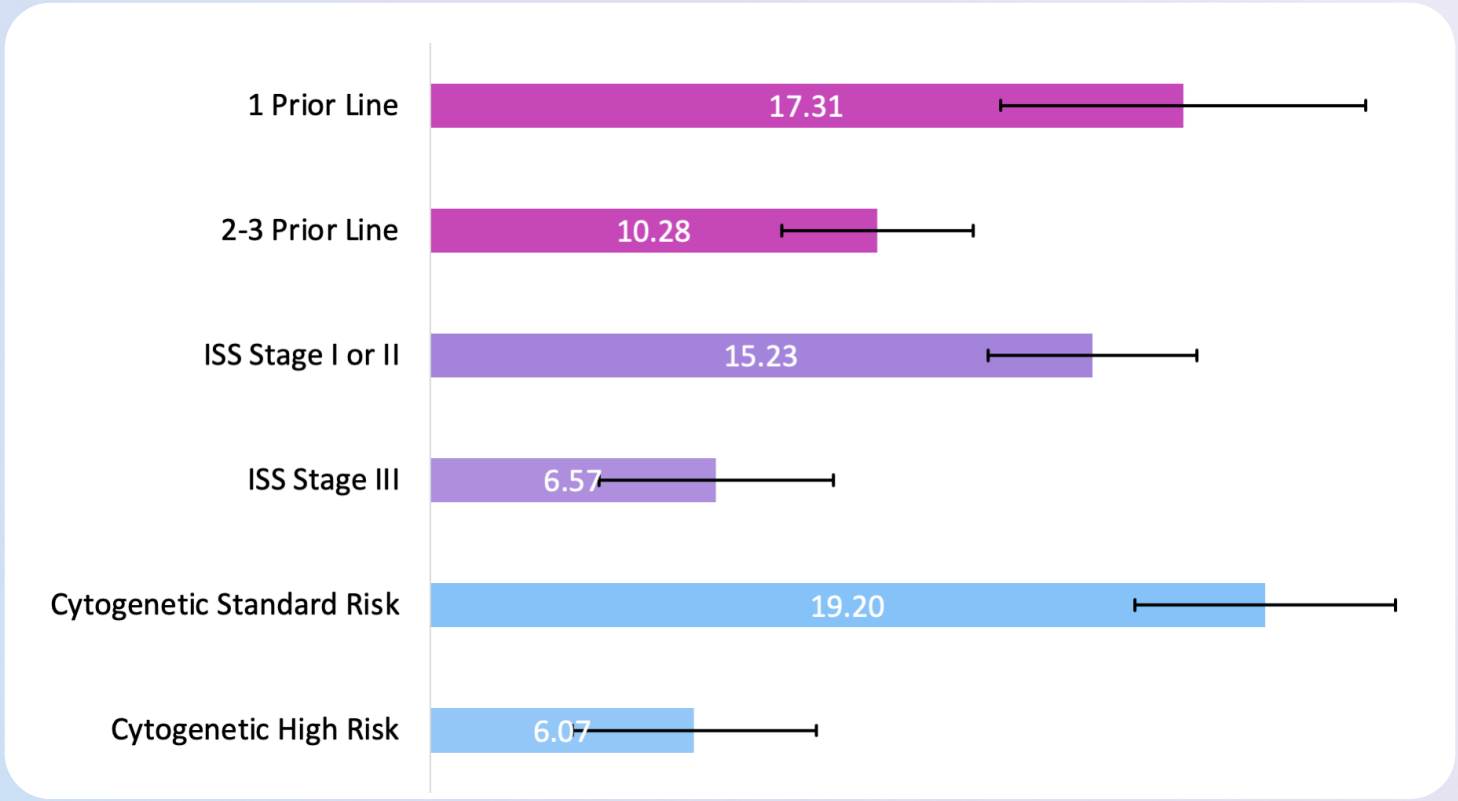
- **Trial design selection**
  AI proposes scenarios; statistician weighs risk & strategy
- **Results interpretation**
  AI generates narrative; human ensures clinical defensibility
- **Regulatory method choice**
  AI references guidelines; human owns submission risk
- **Methods development**
  AI implements papers; human evaluates applicability

IRREPLACEABLE — HUMAN CORE

- **Cross-functional leadership**
  Trust with clinicians, data managers, medical affairs
- **Regulatory accountability**
  Name on submissions; negotiation with FDA/EMA
- **Strategic judgment**
  What to present, what risk to accept, what to defend
- **Mentorship & development**
  Teaching junior biostatisticians; building team judgment
- **Mission & culture**
  Inclusiveness, values, organizational commitment

Use Case 1: Determine the right sample size based on full clinical evidence

Pooled median PFS of early line RRMM by baseline characteristics



80% of clinical trials were designed without conducting systematic review, due to tight timeline and lack of reliable solutions.

Small sample size leads to underpowered study.

Large sample size leads to

- Longer duration of study
- Late to market
- Poor return on investment

Median PFS of early line len-exposed/refractory MM by baseline characteristics

	APPOLO (DPd)	**CARTITUDE-4** (DPd/PVd)	**OPTIMISMM** (PVd)	**CANDOR** (Kd)	**ENDEAVOR** (Kd)	**ARROW** (Kd)
1 Prior Line	14.1	17.4	17.84	21.3	15.6	
2-3 Prior Line	10.7	10.2		12.5	9.7	12.1/8.9
ISS Stage I or II	19.3/12.3	14.7		15.8		
ISS Stage III	6.1	6.2		7.4		
Cytogenetic High Risk	5.8	6.7	8.44	5.6	8.8	
Cytogenetic Standard Risk	21	20.6		16.6		

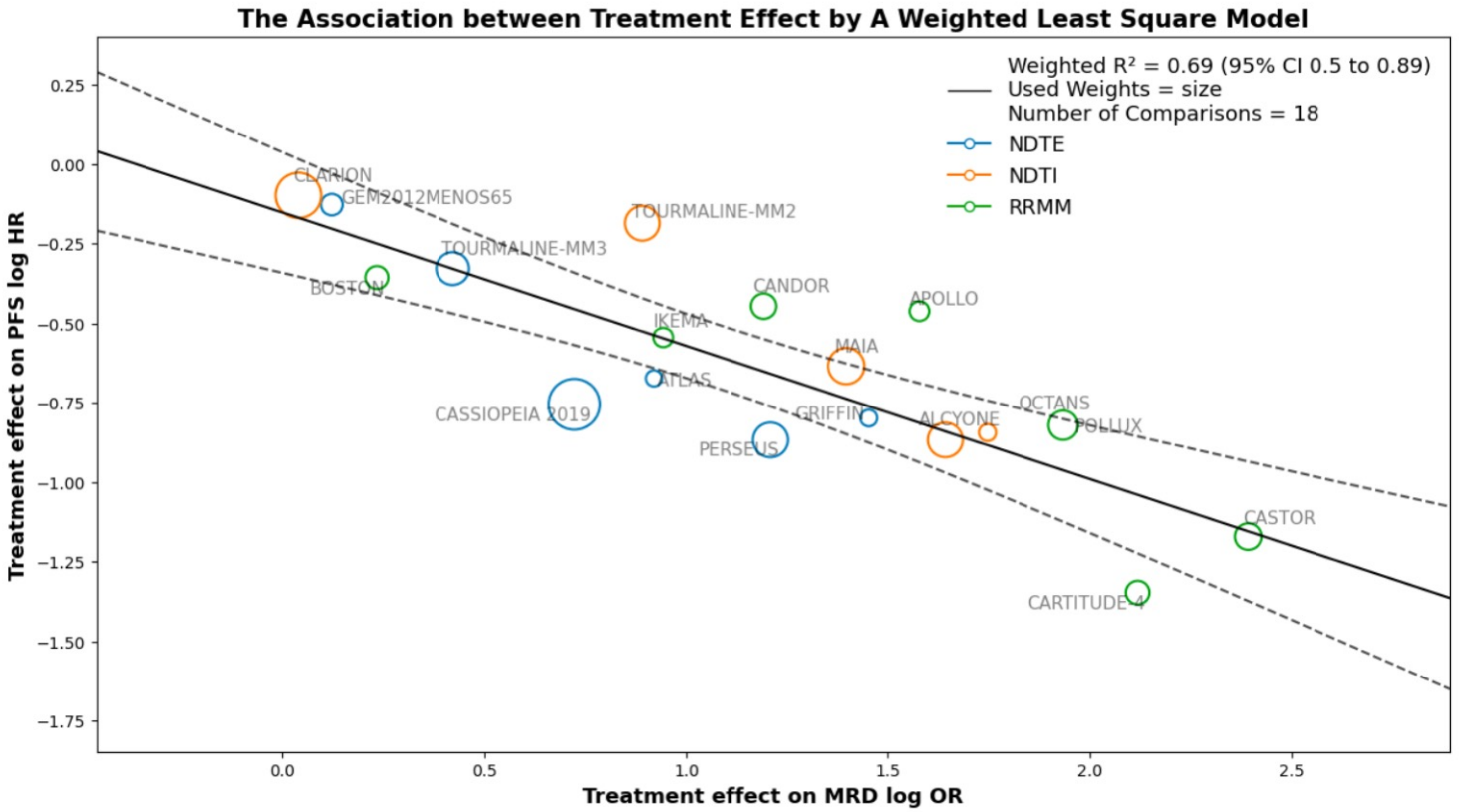
- Clinical trial outcomes by subgroups for Kd, DPd, PVd and EPd for early line lenalidomide-exposed/refractory RRMM were included.
- Data was extracted by AI and curated by human.

Use Case 2: Address FDA's comments on a single arm study in AML with comprehensive evidence

Complete remission (CR) rate of non-targeted therapies for R/R AML patients

Study ID	Study Name	Agent	Population	Sample Size	Median Age (years)	Age Range (years)	CR Rate	CRi Rate	Median OS (months)
NCT02607059	PETHEMA	Venetoclax + HMA	R/R AML	51	68	25-82	10.40%	2.00%	3.4
NCT02421939	ADMIRAL	Salvage chemotherapy	FLT3-mutated R/R AML	124	62	19-85	10.50%	11.30%	5.6
NCT03182244	COMMODORE	Salvage chemotherapy	FLT3-mutated R/R AML (Asian)	118	50		10.20%	9.30%	5
NCT01147939	CLAVELA	Elacytarabine	R/R AML	191	62	22-89	15.00%	8.00%	3.5
NCT01147939	CLAVELA	Investigator Choice	R/R AML	190	63	19-83	12.00%	9.00%	3.3
NCT02152956		Flotetuzumab	R/R AML	50	64	27-82	12.00%		3.2

Use Case 3: Validated surrogacy of MRD negativity in multiple myeloma based on SynthIPD



In partnership with:

MAYO CLINIC

Fred Hutch Cancer Center

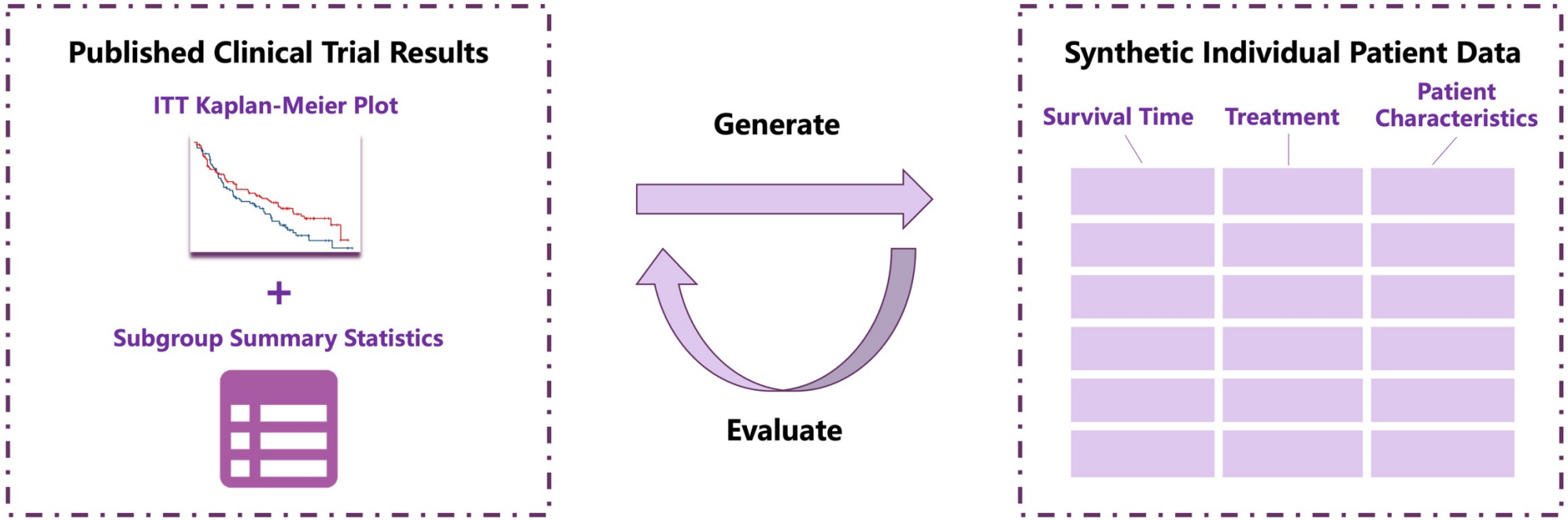
Featured at:

International Myeloma Society

ASCO
AMERICAN SOCIETY OF CLINICAL ONCOLOGY

SynthIPD – Synthetic Individual Patient Data

Reconstructing survival outcome and patient covariate from KM plots and subgroup summary statistics



Published vs. Needed

PUBLISHED

- KM curve + at-risk numbers, by arm
- Median survival + 95% CI, by subgroup
- Hazard ratio + 95% CI, by subgroup
- Subgroup forest plot
- Sometimes: survival rates at landmark times

$\neq$

NEEDED

- One row per patient
- Survival time + event indicator
- Treatment arm
- Covariates (age group, sex, biomarker, …)
- Anything you can subgroup or model on

The fundamental question

Can we generate IPD from published trial reports ***without seeing any real IPD***?

Two gaps in existing reconstruction methods

1 Pixel-based digitization is noisy

Existing tools (WebPlotDigitizer + IPDfromKM) require a human to click on every drop and tick on a screenshot of the KM plot.

0-20% maximum relative error in reconstructed medians, depending on the manual efforts.

2 No covariates

You do not get age, sex, biomarker status, or any of the subgroup variables the trial actually reported on.

Cannot re-run subgroup analyses, do covariate-adjusted comparisons, or pool patients by biomarker across trials.

SynthIPD in 3 steps

STEP 1

DIGITIZE

Extract event/censor coordinates from the KM plot using its SVG representation.

→ *survival times, status, arm*

STEP 2

Cov-Generation

Search for a covariate assignment that reproduces the published subgroup table.

→ *covariate vector $\tilde{x}_i$ for each patient*

STEP 3

Combine

Concatenate step 1 and step 2 into a synthetic IPD table.

→ *one row per patient, ready to analyze*

What you only need from the publication

KM plot (vector PDF) | **At-risk table** | **Subgroup median + 95% CI** | **Subgroup HR + 95% CI** | **Censor counts (preferred)**

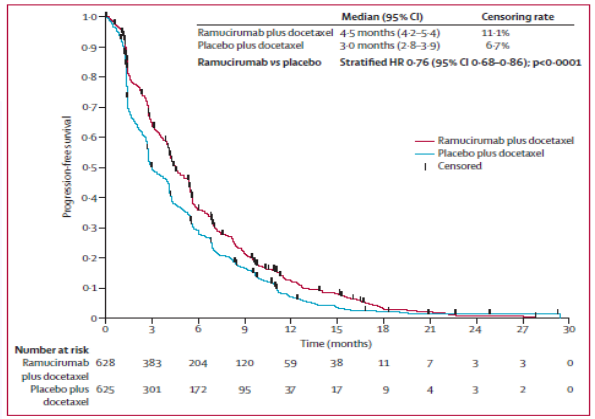
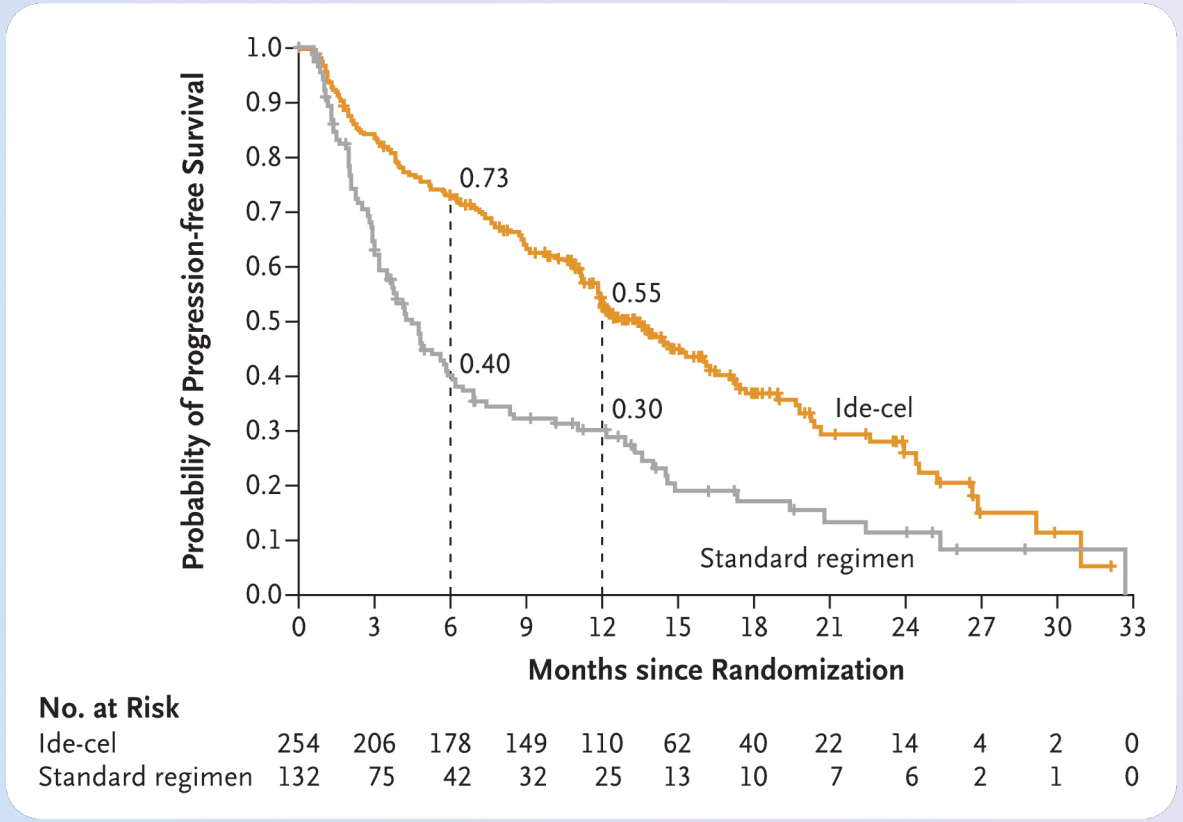


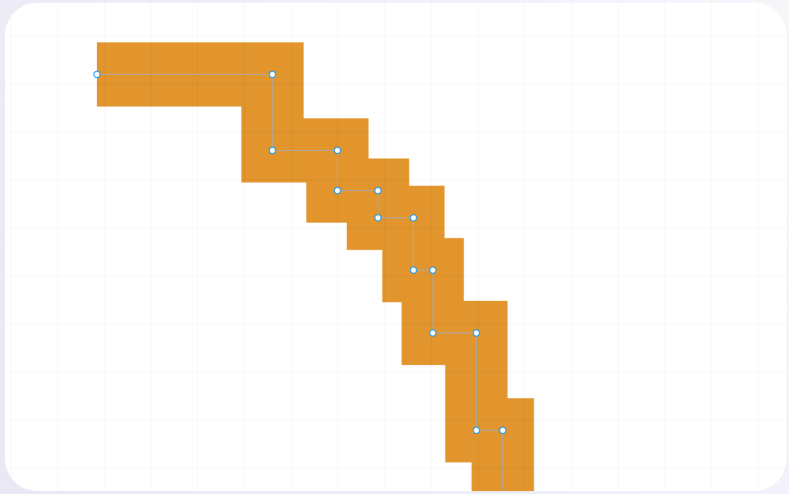
Figure 3: Kaplan-Meier estimates of progression-free survival in the intention-to-treat population
HR=hazard ratio.

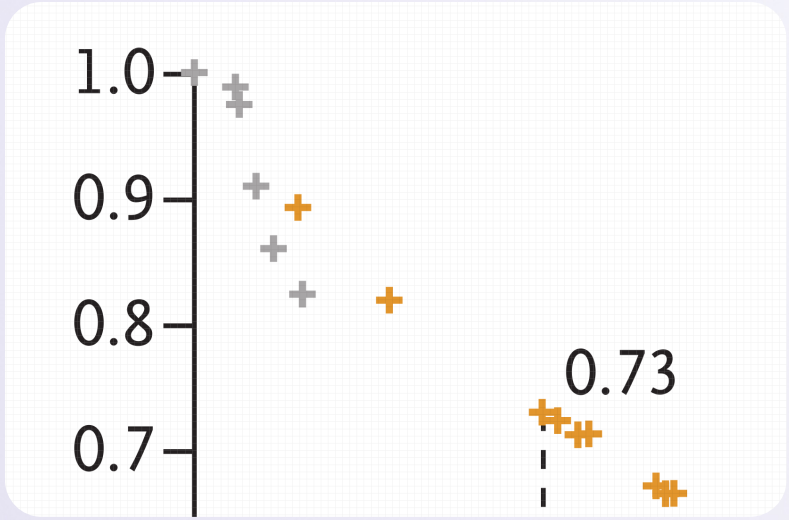
	Events/patients		Median (95% CI) progression-free survival, months		Hazard ratio (95% CI)
	Daratumumab plus pomalidomide and dexamethasone group	Pomalidomide and dexamethasone group	Daratumumab plus pomalidomide and dexamethasone group	Pomalidomide and dexamethasone group	
Sex					
Male	46/79	54/82	10·7 (7·4–19·3)	7·2 (4·9–10·6)	0·69 (0·47–1·03)
Female	38/72	52/71	15·0 (8·2–NE)	6·5 (4·7–9·3)	0·54 (0·35–0·82)
Age (years)					
<65	36/63	41/60	9·2 (4·6–21·0)	5·8 (3·7–12·6)	0·69 (0·44–1·09)
≥65	48/88	65/93	14·2 (9·9–NE)	7·0 (6·1–10·1)	0·55 (0·38–0·81)

Digitization of the KM plot at pixel-level accuracy

Enables reconstruction of the exact survival time and censor status of each patient in the ITT population





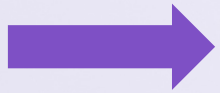


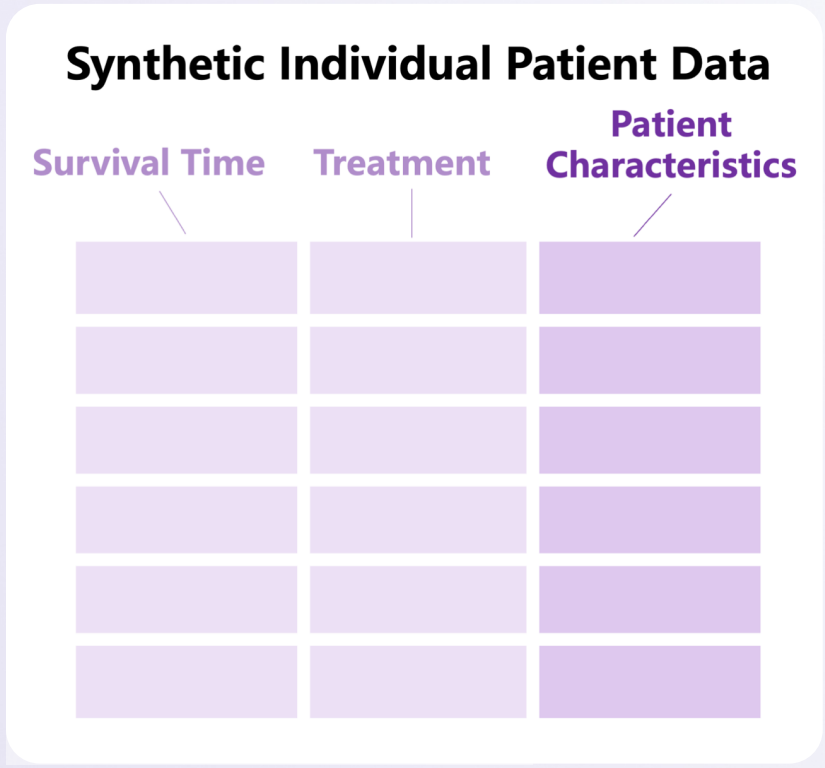
Simulation of covariates to match the subgroup results

A novel iterative algorithm with substantial computing power was used to simulate patient characteristics

Figure S1. Subgroup Analysis of Progression-free Survival.

	Ide-cel	Standard Regimen	Hazard Ratio (95% CI)	Ide-cel	Standard Regimen
	No. of progression or death events/total no.			*Median progression-free survival — months (95% CI)*	
All patients	149/254	93/132	0.51 (0.39-0.66)	13.3 (11.8-16.1)	4.4 (3.4-5.9)
Age group					
<65	93/150	51/78	0.57 (0.40-0.80)	12.7 (11.1-17.2)	4.4 (3.2-5.9)
≥65	56/104	42/54	0.43 (0.28-0.64)	13.3 (11.2-16.6)	4.7 (3.0-6.9)
Age group					
<65	93/150	51/78	0.57 (0.40-0.80)	12.7 (11.1-17.2)	4.4 (3.2-5.9)
65-74	49/92	36/45	0.42 (0.27-0.65)	13.3 (11.2-17.4)	4.2 (3.0-7.5)
75-84	7/12	6/9	0.59 (0.19-1.78)	13.2 (4.0-NE)	5.7 (0.9-NE)
Region					
North America	84/144	60/82	0.50 (0.36-0.70)	13.8 (11.3-19.0)	4.7 (3.2-6.1)
Europe	63/106	32/45	0.44 (0.28-0.68)	12.4 (10.6-16.1)	3.9 (2.1-6.5)
Japan	2/4	1/5	NC	NE (2.0-NE)	20.7 (NE-NE)
Sex					
Male	92/156	55/79	0.53 (0.38-0.75)	13.8 (11.3-17.2)	4.8 (3.5-7.5)
Female	57/98	38/53	0.47 (0.31-0.71)	12.0 (10.2-17.7)	3.7 (2.8-5.8)
Race					
White	101/172	54/78	0.52 (0.37-0.73)	13.6 (11.3-17.2)	4.9 (3.0-8.4)
Non-white	14/28	18/27	0.59 (0.29-1.22)	20.3 (11.1-24.0)	6.9 (3.4-20.7)
Race					
White	101/172	54/78	0.52 (0.37-0.73)	13.6 (11.3-17.2)	4.9 (3.0-8.4)
Asian	4/7	1/5	NC	11.1 (0.8-NE)	20.7 (NE-NE)
African American	8/18	13/18	0.50 (0.20-1.23)	20.3 (8.9-NE)	6.9 (3.7-22.5)
Other	2/3	4/4	NC	17.7 (12.7-NE)	2.7 (1.6-NE)
Ethnicity					
Hispanic or Latino	5/11	5/8	0.21 (0.05-0.93)	22.6 (1.3-NE)	6.1 (1.0-NE)
Not Hispanic or Latino	109/188	68/98	0.56 (0.41-0.76)	13.8 (11.3-16.2)	4.9 (3.5-7.5)
Anti-CD38 class refractory					
Yes	143/242	88/124	0.51 (0.39-0.67)	12.5 (11.3-15.1)	4.4 (3.2-5.9)
No	6/12	5/8	0.40 (0.11-1.40)	19.0 (3.0-NE)	5.3 (0.6-NE)
Daratumumab refractory					
Yes	143/242	88/123	0.51 (0.39-0.67)	12.5 (11.3-15.1)	4.4 (3.2-5.9)
No	6/12	5/9	0.40 (0.11-1.40)	19.0 (3.0-NE)	5.3 (0.6-NE)
Double-class (IMiD agent and PI) refractory					
Yes	106/169	72/91	0.47 (0.34-0.63)	11.2 (8.7-12.7)	3.5 (2.9-4.7)
No	43/85	21/41	0.65 (0.38-1.11)	17.3 (14.5-22.6)	12.9 (4.3-NE)
Triple-class refractory*					
Yes	103/164	70/89	0.46 (0.34-0.62)	11.2 (8.7-12.5)	3.5 (2.9-4.7)
No	46/90	23/43	0.65 (0.39-1.09)	17.3 (14.5-22.6)	12.9 (4.3-22.5)
Penta-refractory†					
Yes	12/15	3/5	0.63 (0.17-2.33)	3.9 (2.0-12.0)	3.0 (1.0-NE)
No	137/239	90/127	0.49 (0.37-0.64)	13.8 (11.9-16.6)	4.7 (3.4-5.9)
Revised ISS stage at baseline					
I or II	113/200	78/108	0.48 (0.36-0.64)	14.7 (12.1-17.3)	4.4 (3.2-5.9)
III	27/31	8/14	0.86 (0.39-1.92)	5.2 (1.8-7.2)	3.0 (0.8-6.1)
Tumor burden‡					
<50%	99/172	60/90	0.47 (0.34-0.65)	14.1 (12.0-17.2)	4.2 (3.2-6.5)
≥50%	44/71	28/34	0.60 (0.37-0.97)	11.0 (7.2-16.2)	4.9 (2.3-10.1)
Extramedullary plasmacytoma					
Yes	48/61	28/32	0.40 (0.25-0.65)	7.2 (4.0-11.8)	2.0 (1.3-3.0)
No	100/192	65/100	0.51 (0.37-0.70)	16.1 (12.7-19.8)	5.9 (4.3-10.1)
Number of prior anti-myeloma regimens					
2	41/78	26/39	0.51 (0.31-0.84)	15.1 (12.7-19.7)	4.8 (3.2-13.3)
3 or 4	108/176	67/93	0.51 (0.37-0.69)	12.0 (10.2-14.7)	4.2 (3.0-5.9)





N9741 — colorectal cancer trial, blind test against Mayo Clinic

THE TRIAL

N9741 (Goldberg 2002; Sanoff 2008): phase 3 in metastatic colorectal cancer.

FOLFOX (n = 421) vs. **IROX** (n = 383). Primary endpoint: progression-free survival.

True IPD: not publicly available. Mayo Clinic holds the data.

THE BLIND COMPARISON

Test design

1. Run SynthIPD on the published inputs only.
2. Send the synthetic IPD back to Mayo.
3. Mayo overlays it against their held-out KM truth.
4. Compare summary statistics and subgroup KM curves.

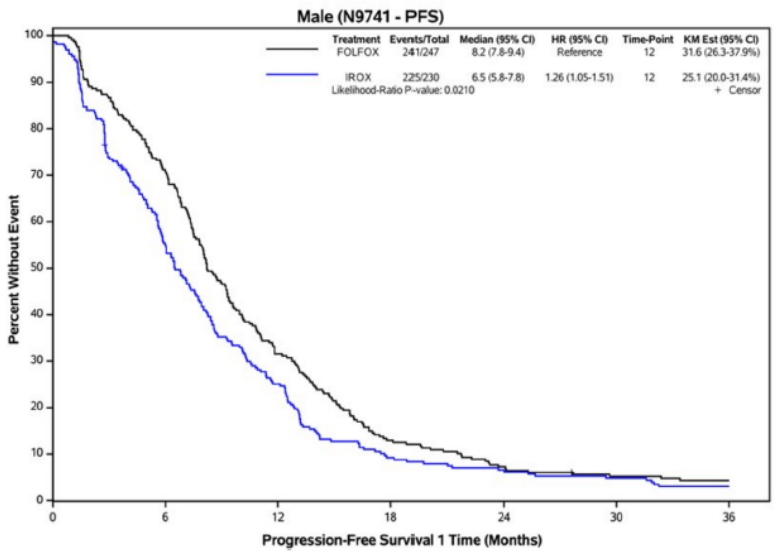
INPUTS WE GOT

- KM plot (ITT, both arms)
- At-risk table
- Subgroup table by Gender

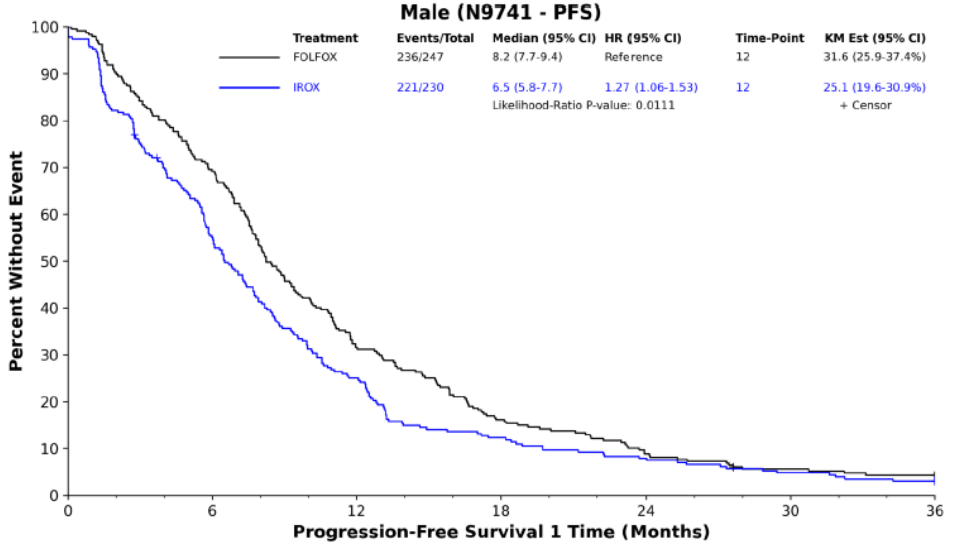
SynthIPD reproduces the subgroup table to reported precision

						HR (95% CI)
M						26 (1.05–1.51)
Fer						29 (1.04–1.61)
M						26 (1.05–1.51)
Fer						29 (1.04–1.61)

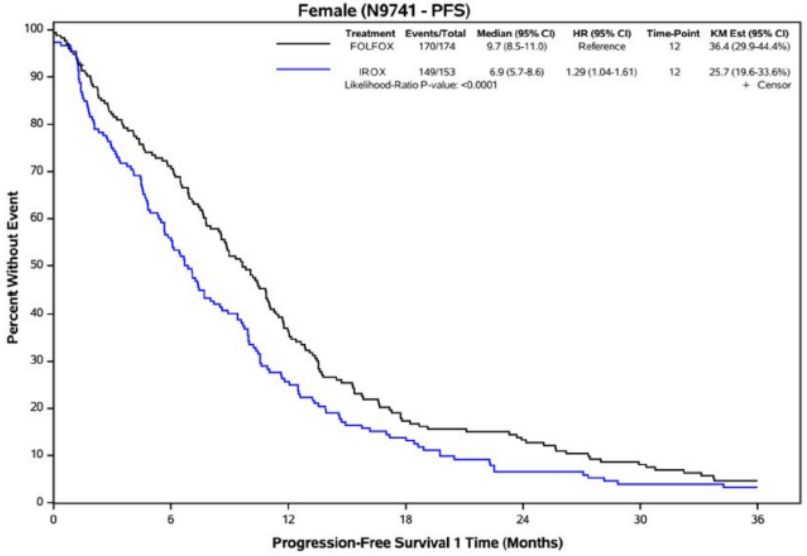
(a) Mayo Clinic, Male



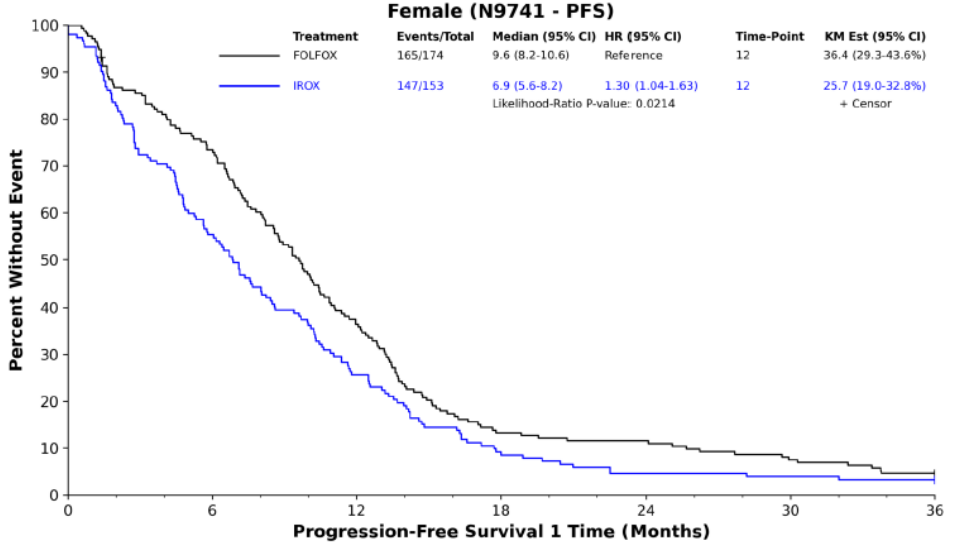
(b) SynthIPD, Male



(c) Mayo Clinic, Female



(d) SynthIPD, Female



Clinical evidence that wasn't in the original paper

Insight	Arm	Subgroup	Estimate (95% CI)		
			12 months	24 months	36 months
Survival rate	FOLFOX	Female	0.36 (0.29, 0.44)	0.11 (0.08, 0.18)	0.05 (0.02, 0.09)
		Male	0.32 (0.26, 0.38)	0.09 (0.06, 0.13)	0.04 (0.02, 0.08)
	IROX	Female	0.26 (0.20, 0.34)	0.05 (0.02, 0.10)	0.03 (0.01, 0.08)
		Male	0.25 (0.20, 0.31)	0.08 (0.06, 0.12)	0.03 (0.01, 0.06)
RMST (months)	FOLFOX	Female	8.33 (7.81, 8.99)	10.42 (9.45, 11.48)	11.33 (10.01, 12.63)
		Male	8.02 (7.63, 8.52)	10.37 (9.41, 11.16)	10.94 (9.91, 12.12)
	IROX	Female	7.65 (7.10, 8.32)	9.21 (8.10, 10.35)	9.81 (8.50, 11.04)
		Male	7.40 (6.95, 7.92)	9.10 (8.20, 10.05)	9.65 (8.45, 10.78)

Where SynthIPD can plug into adaptive trial design

Use case 1

Control-arm benchmarks for sample size

Covariate-adaptive randomization (CAR) and CARA designs depend on assumed control-arm performance. Pool synthetic IPD from 5-10 historical control arms to get an empirically grounded estimate for the control arms.

Use case 2

Indirect matching comparison

Two trials may have different patient populations. The SynthIPD can be used to compare the result of two trials without seeing the true IPD data.

Use case 3

Surrogacy and endpoint association

Modern adaptive trials increasingly use surrogate endpoints. SynthIPD-generated patient-level data enables association analyses between, e.g., MRD and PFS across trials (Ren et al., CRC 2026).

Use case 3

Priors for Bayesian adaptive trials

Bayesian adaptive designs need informative priors on control-arm parameters. SynthIPD turns a stack of published curves into a usable prior distribution.

HopeAI

Bring hope to patients through AI.