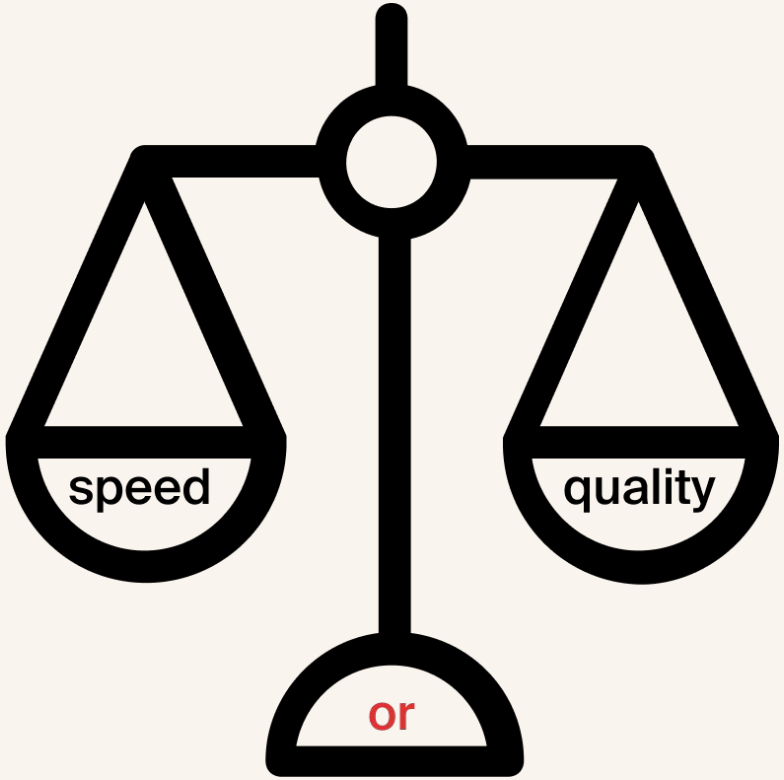
AI Got the Power: Streamlining Clinical Data Creation

Presenters: Emily Yates, Formation Bio, USA

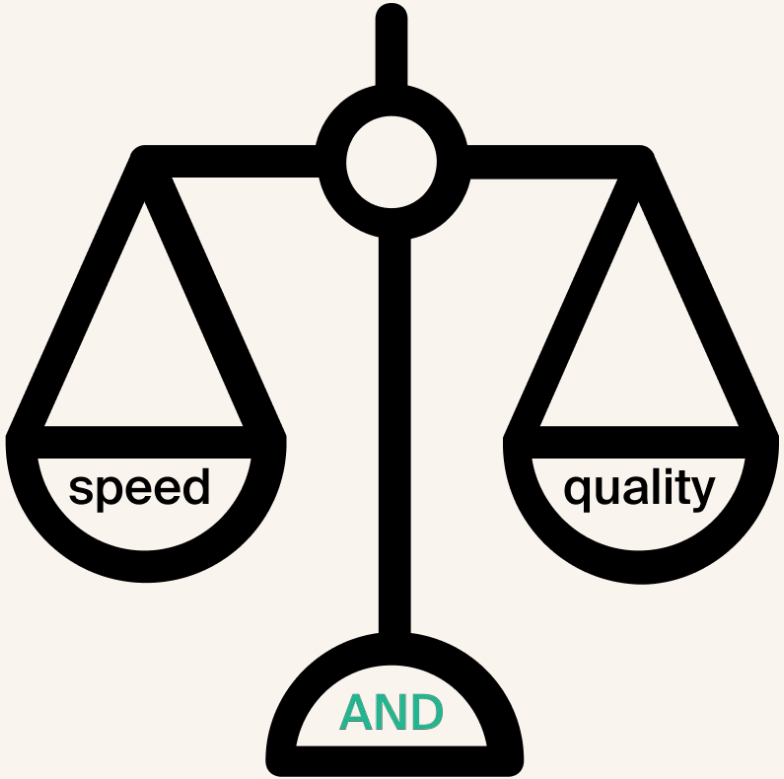
Co-Author: Andrew Burd & Matt Luppino, Formation Bio, USA

Paper ML14

When innovating, we balance...



When using AI for test data generation...



The process for creating test data is highly manual

Blocked until many key study documents are finalized

Manually entered into the EDC or vendor system

8-10 subjects at a single site

The process for creating test data is inadequate & has consequences

Blocked until many key study documents are finalized

Manually entered into the EDC or vendor system

8-10 subjects at a single site

Delays start of programming work

Time and resource-intensive

Inadequate coverage of edge cases leading to errors in programming

AI can unlock the untapped potential value of test data

Shift-left on programming work to deliver more & on tighter deadlines

Save SSU time and reduce cost by eliminating manual work

Mitigate risks and prevent future issues through robust testing

+ new capabilities

- Enhanced medical and safety monitoring
- Protocol deviation management
- Vendor data reconciliation
- Early TLFs

Comparing AI Solutions

A good AI tool creates high quality data quickly

High Quality Data

- Clean / “perfect” data
- Realistic errors in the data
- Structure matches the source system

Scalability

- High volume of data
- Repeatable process
- Limited human interaction

Easy to use

- Quickly to configure
- Widely available
- Low learning curve

Comparative analysis of AI test data generation solutions

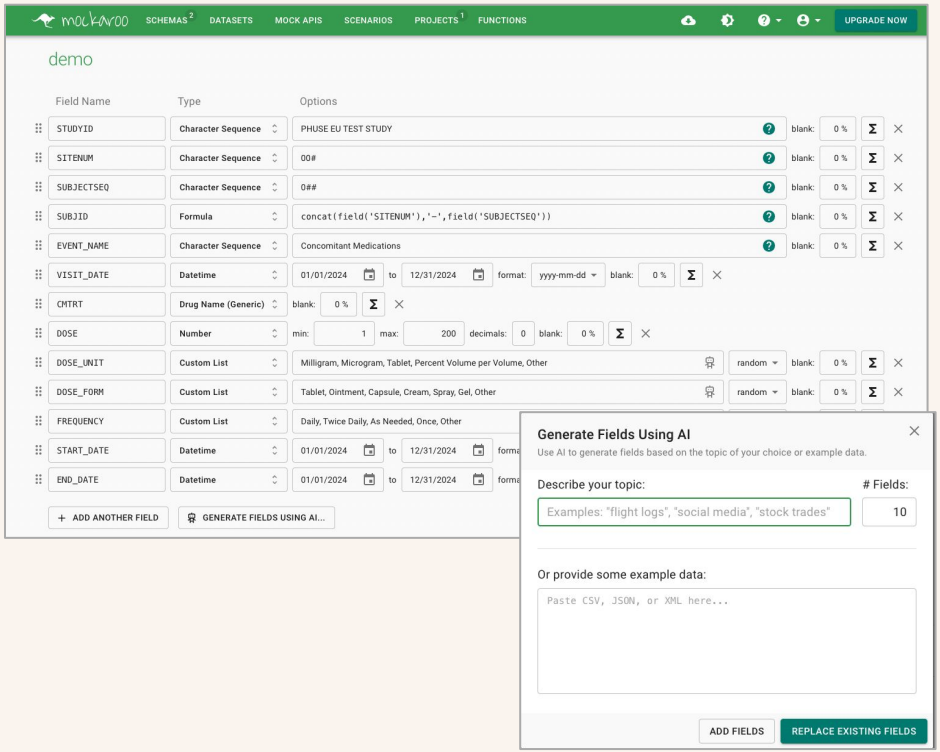
A framework to find the best AI solution

01 Mockaroo

02 ChatGPT

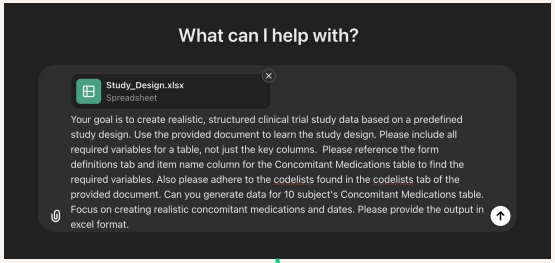
03 Integrated LLM + EDC solution

Mock data generation with Mockaroo



CRITERIA	MOCKAROO	RATING
High Quality Data	• Able to have clean data • Limited error simulation • Limited ability to match structure of source system	Moderate
Scalable	• Generates large volume of data • Repeatable process • Limited human interaction	High
Easy to Use	• Configuration takes a while • Free and available • Medium learning curve	Moderate

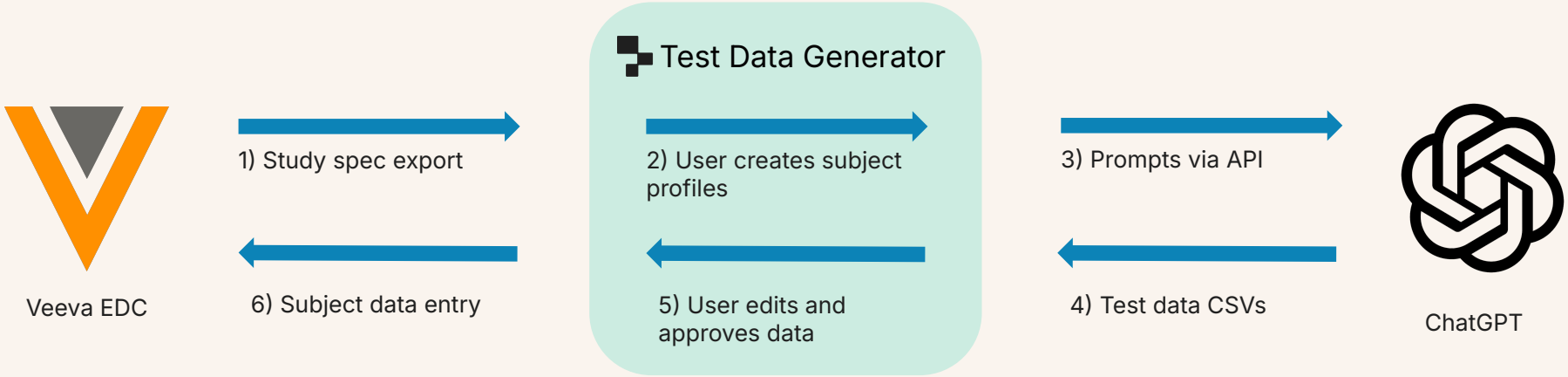
Mock data generation with ChatGPT



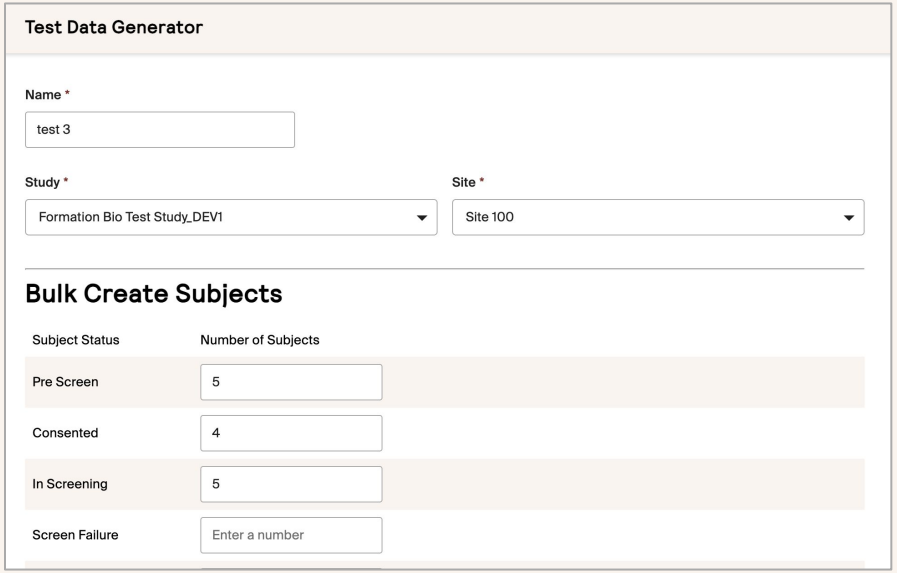
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Subject ID	CMTRT	CMINDC	M_CMSTDAT	M_CMSTTIM	CMPRIOR	CMDOSE	CMDOSTOT	CMDOSU	CMDOSFRM	CMDOSFRQ	CMROUTE	M_CMON
2	SUBJ_1	Metformin	Pain management	2022-03-06	00:00	No	50		ml	Capsule	As needed	Subcutaneous	No
3	SUBJ_1	Atorvastatin	Diabetes	2020-03-23	00:00	No	76	83	ml	Tablet	Once daily	Intravenous	Yes
4	SUBJ_2	Aspirin	Diabetes	2020-02-03	00:00	No	13	119	ml	Injection	Twice daily	Subcutaneous	Yes
5	SUBJ_3	Lisinopril	Hypertension	2022-08-02	00:00	Yes	64	103	g	Tablet	As needed	Intravenous	No
6	SUBJ_3	Atorvastatin	Pain management	2023-01-21	00:00	Yes	58		g	Tablet	As needed	Subcutaneous	Yes
7	SUBJ_3	Atorvastatin	Diabetes	2021-11-02	00:00	Yes	85		g	Capsule	Twice daily	Subcutaneous	Yes
8	SUBJ_4	Aspirin	Hypertension	2023-12-17	00:00	No	90		g	Injection	Twice daily	Oral	No
9	SUBJ_4	Atorvastatin	Pain management	2021-11-05	00:00	Yes	30	58	g	Capsule	Once daily	Oral	No
10	SUBJ_4	Ibuprofen	Hypertension	2023-03-22	00:00	Yes	53	135	mg	Tablet	As needed	Intravenous	No
11	SUBJ_5	Atorvastatin	Hypertension	2021-04-03	00:00	Yes	96		mg	Injection	As needed	Subcutaneous	No
12	SUBJ_5	Atorvastatin	Pain management	2022-09-03	00:00	Yes	32		ml	Injection	Once daily	Oral	Yes
13	SUBJ_6	Metformin	Pain management	2023-06-18	00:00	Yes	43		mg	Tablet	Twice daily	Oral	No
14	SUBJ_6	Lisinopril	Diabetes	2023-01-15	00:00	Yes	54	72	mg	Tablet	As needed	Oral	Yes
15	SUBJ_6	Metformin	Diabetes	2020-07-21	00:00	No	34	89	ml	Injection	Twice daily	Oral	Yes
16	SUBJ_7	Aspirin	Diabetes	2020-06-15	00:00	Yes	80	125	g	Capsule	As needed	Oral	No
17	SUBJ_7	Ibuprofen	Diabetes	2024-08-03	00:00	No	44		mg	Injection	Once daily	Oral	Yes
18	SUBJ_7	Aspirin	Diabetes	2020-09-06	00:00	Yes	100		mg	Injection	Twice daily	Subcutaneous	Yes
19	SUBJ_8	Metformin	Diabetes	2024-01-12	00:00	No	54		ml	Injection	Once daily	Intravenous	No
20	SUBJ_9	Ibuprofen	Diabetes	2022-08-17	00:00	No	47	154	ml	Tablet	Once daily	Oral	No
21	SUBJ_10	Ibuprofen	Hypertension	2022-04-21	00:00	Yes	90	197	ml	Injection	As needed	Subcutaneous	No
22	SUBJ_10	Metformin	Hypertension	2020-04-07	00:00	No	40	63	ml	Capsule	As needed	Oral	Yes
23	SUBJ_10	Atorvastatin	Pain management	2023-10-28	00:00	No	99		ml	Capsule	Once daily	Intravenous	No

CRITERIA	CHATGPT	RATING
High Quality Data	• Able to have clean data • Able to have error simulation • Limited ability to match structure of source system	Moderate
Scalable	• Generates large volume of data • Repeatable process • Some human interaction needed for query refinement	High
Easy to Use	• No configuration needed • Free and available • Medium learning curve	High

Integrated LLM + EDC solutions



Integrated LLM + EDC solutions



CRITERIA	INTEGRATED LLM + EDC	RATING
High Quality Data	• Able to have clean data • Able to have error simulation • Matches structure of source system	High
Scalable	• Generates large volume of data • Repeatable process • Human interaction needed	High
Easy to Use	• No configuration needed • Engineering investment needed • Medium learning curve	Moderate

There's no perfect solution, but AI creates good options

- AI tools are good at solving for scalability
- We can achieve high quality and speed at the cost of complexity
- Even moderate data quality is enough to jump start programming and innovation
- Learning curve for AI can be tackled through trainings & hackathons to make the tools easier to use

CRITERIA	MOCKAROO	CHATGPT	INTEGRATED LLM + EDC
High Quality Data	Moderate	Moderate	High
Scalable	High	High	High
Easy to Use	Moderate	High	Moderate

Formation Bio

Questions?