

EP 6

August 24, 2025 · 02:05:32

Explainable Medical AI - Dr. Rich Caruana

Dr. Rich Caruana — Principal Researcher, Microsoft Research



Dr. Rich Caruana on models a doctor can read: explainable boosting machines that match black boxes on patient data, what their graphs reveal about care, and why accurate is not correct.

50,000

lives a year earlier dialysis might save, if it holds up
[25:53](#)

85

the age at which pneumonia care turned less aggressive
[17:56](#)

4 + 1

doctors and a language model, reading the same 50 graphs
[40:52](#)

100+

healthcare papers that already use these models
[57:11](#)

KEY TAKEAWAYS

1 Accurate is not the same as correct

A model scores well by repeating its data's quirks. It can learn that very high creatinine is safe, because those patients got dialysis. [43:12](#)

2 A glass box can match a black box

On tables of numbers, explainable boosting machines are as accurate as deep nets and boosted trees, and every effect is a graph. [7:58](#)

3 The graphs show how care is given

In the pneumonia data, risk jumps at 85 and falls at 100: the very old were treated less aggressively, centenarians more. [17:56](#)

4 Round numbers set thresholds

Treatment starts at 50, 100, 180. Giving dialysis at 85 instead of 100 might save lives, which only a trial could confirm. [25:14](#)

5 Let a language model read first

Given the graphs as tables, an LLM found nearly every surprise the doctors found, and some false ones. It tells clinicians where to look. [40:52](#)

“There's nothing more dangerous in life and in science than a false assumption.”

Dr. Rich Caruana · [1:48:23](#)

CHAPTERS



0:00	Intro	2 min	40:26	Four doctors and an LLM	3 min
1:42	Why interpretability matters	2 min	43:12	Accuracy is not correctness	7 min
3:15	A short history of machine learning	3 min	50:40	Where it fits in a hospital	10 min
6:30	Glass-box models for tabular data	2 min	1:00:16	Connecting EBMs to LLMs	10 min
8:20	What an EBM is	7 min	1:10:27	Intelligible, the company	11 min
14:52	Pneumonia risk by age	8 min	1:21:34	Who the customers are	5 min
23:09	Treatment effects in blood tests	6 min	1:26:59	Keeping the detail in the graphs	6 min
29:26	Blood pressure and round numbers	2 min	1:33:26	Between labs and universities	8 min
31:41	Risk in childbirth	5 min	1:41:38	Advice for young scientists	9 min
36:30	LLMs that read the graphs	4 min	1:50:43	Healthcare in ten years	15 min

TERMS EXPLAINED

Glass-box model

A model whose reasoning can be read directly, unlike a black box.

Explainable boosting machine

A generalized additive model trained with boosted trees: the EBM of this episode.

Treatment effect

A change in outcomes caused by the care patients received, not by their underlying health.

Generalized additive model

A model that adds up one curve per input, so each effect can be drawn as a graph.

Tabular data

Data in rows and columns of numbers, like a patient's test results.

Multitask learning

Training one model on several related tasks so that each helps the others.

MENTIONED IN THIS EPISODE

InterpretML

Intelligible

Intelligible Models for HealthCare

LLMs and glass-box models (2023)

Generalized additive models

Listen, watch or download the transcript: aarsentyev.com/podcast/ep6-rich-caruana

Each time opens the video at that moment. Hosted by Arseniy Arsenyev. A summary of a conversation, not medical advice.