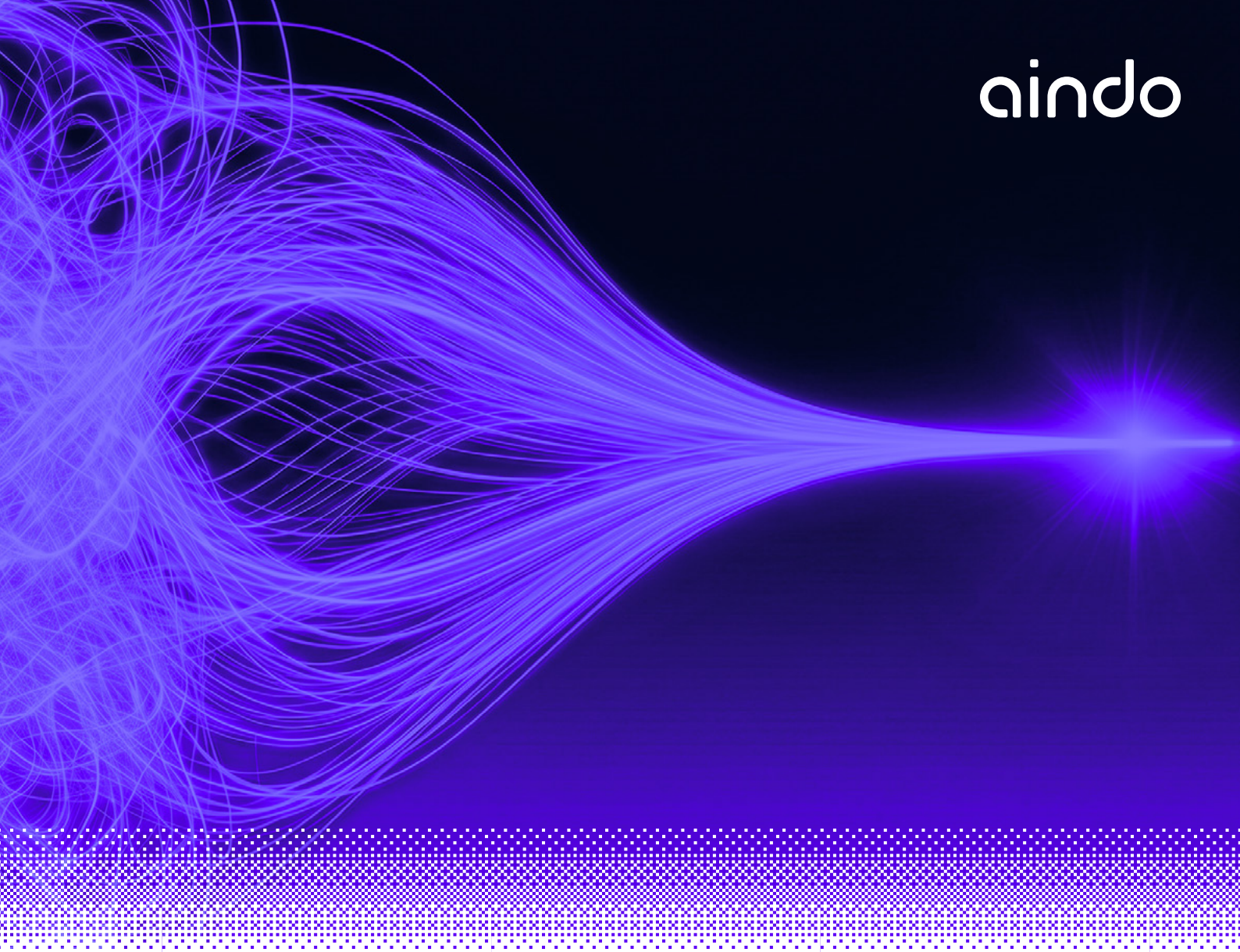


Use Case: Predicting Missing Clinical Follow-Ups with AI-Generated Synthetic Data

aindo

Extending the Evidence Timeline: Predicting Missing Clinical Follow-Ups with AI-Generated Synthetic Data

How a certified, privacy-by-design generative AI model reconstructed missing patient follow-up data in cardiovascular prevention — and outperformed standard machine learning in the process.

Index

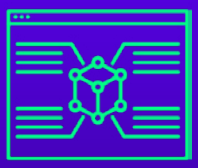
Executive Summary

Real-world evidence generation in cardiovascular prevention is often held back not by a shortage of patients, but by incomplete follow-up: visits are missed, spaced irregularly, or simply haven't happened yet within the window available for analysis. Working with a global biopharmaceutical company, Aindo applied its Europrivacy-certified generative AI technology to generate — impute — these missing follow-up laboratory measurements, and then used those generated values to predict whether each patient reached their LDL-C target, validating that this two-step approach showed a consistent advantage over a leading machine learning benchmark across all cross-validation folds.

Key Results at a Glance

Predictive accuracy

using the AI-generated (imputed) follow-up lab values, the model correctly classified patients as having reached their LDL-C target or not, 77% of the time under a conservative goal and 74% of the time under a stricter one — a consistent advantage across all cross-validation folds over both an uninformed statistical guess (59% / 52%) and gradient boosting, a leading machine learning method (72% / 71%).



Data fidelity

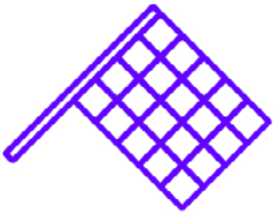
Across every categorical variable in the dataset, the statistical difference between the synthetic and real data was near zero — a result that would typically be considered an excellent match even between two samples drawn from the same real population.

Data efficiency

All of this was achieved from a dataset of only 700 patients, a sample size that would normally be considered too small for robust machine learning.

The Challenge

In secondary cardiovascular prevention, whether a patient reaches their LDL-C target shapes major downstream decisions — therapy escalation, clinical guidance, reimbursement. But the real-world data needed to study this is rarely complete. Follow-up visits are missed, irregularly spaced, or simply haven't occurred yet within the observation window a research team has access to. The conventional fix — wait for more data to accumulate — can add months or years to a research timeline, and that delay translates directly into slower evidence, slower decisions, and slower patient access to appropriate care.



The Approach

Aindo, working with a global biopharmaceutical company, applied its proprietary generative data synthesis technology in two sequential steps. First, the model generated — imputed — statistically faithful estimates of the follow-up laboratory measurements each patient's own clinical trajectory would be expected to produce, rather than fabricating implausible values or simply waiting for real data to arrive. Second, those generated values were used to classify whether each patient had reached their LDL-C target, turning a data-imputation problem into a downstream outcome-prediction task.

The privacy safeguards aren't a claim made about the technology; they are independently audited and certified.

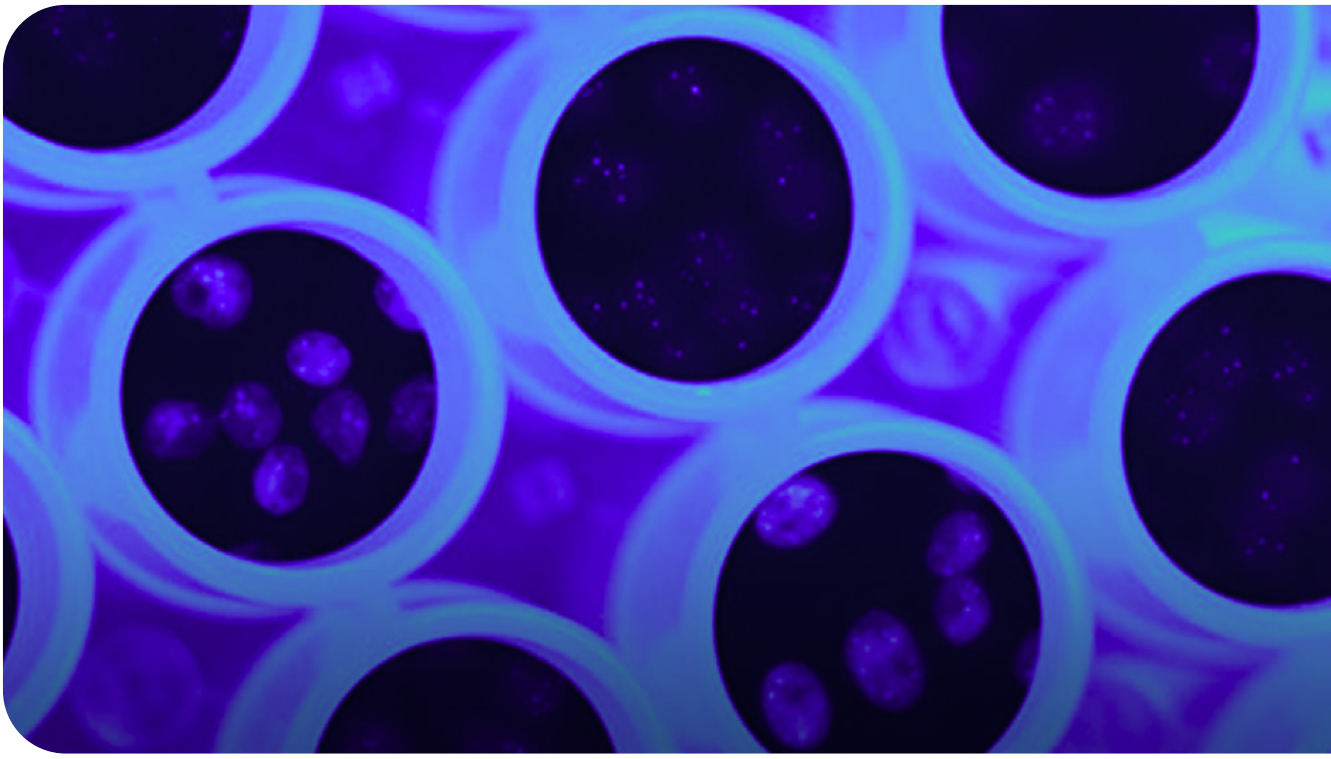
The technology underlying this approach holds Europrivacy certification, the GDPR European Data Protection Seal officially recognized under Article 42 GDPR across all EU and EEA member states. Aindo was the first company in Europe to obtain this certification specifically for the generation of synthetic healthcare data, and the certification's scope extends to the use of generative models for predictive purposes on clinical data — precisely the application demonstrated here. In practice, this means the privacy safeguards aren't a claim made about the technology; they are independently audited and certified.

Starting from a real-world dataset of 700 patients in a cardiovascular disease area — 72 clinical variables spanning up to two decades of practice, including a range of modern treatment regimens — the model was trained to generate (impute) follow-up LDL-C, HDL-C, triglycerides, and total cholesterol from each patient's baseline profile and treatment intensity.

How the Model work

The model is a generative neural network: rather than producing a single fixed answer, it learns the statistical patterns connecting a patient's baseline profile and treatment to their likely future lab values, and can then generate multiple plausible outcomes for the same patient — capturing not just an expected value, but the uncertainty around it. When a real-world clinical dataset is too small, on its own, for a model to learn reliably — as was the case here — Aindo's approach can first train on real data enriched with additional synthetic samples generated with the help of a large language model, then fine-tune on the real data alone. Aindo applies this enrichment step selectively, depending on how much real data is available; here, given the dataset's limited size, it was used to help the model converge on realistic patterns. The values generated at this stage are the imputed lab measurements; the separate classification step described earlier then uses them to predict whether each patient reached their LDL-C target.

Capturing not just an expected value, but the uncertainty around it. When a real-world clinical dataset is too small, on its own, for a model to learn reliably



Validation: What the Numbers Mean

Data fidelity was assessed using the Jensen-Shannon divergence, a standard statistical measure of how different two distributions are

Two independent questions were tested: does the synthetic data itself statistically resemble the real data, and do the model's predictions actually work?

Data fidelity was assessed using the Jensen-Shannon divergence, a standard statistical measure of how different two distributions are (0 means identical, and higher values mean increasingly different). Across every categorical variable in the dataset, this value stayed below 0.008 — indicating the synthetic data's patterns are, for practical purposes, statistically indistinguishable from the real cohort's.

Predictive performance was assessed using AUC (Area Under the ROC Curve), the standard way to measure how well a model distinguishes between two outcomes — here, patients who reached their LDL-C target versus those who didn't. An AUC of 0.50 means the model is no better than a coin flip; 1.00 means perfect discrimination. The results, validated using 5-fold cross-validation (the data was split five different ways, with the model tested on each portion after training on the rest, to confirm the result wasn't a fluke of one particular split):

MODE	AUC / Conservative goal	AUC / Strict goal
Uninformed statistical estimate	0.59	0.52
Gradient boosting (leading ML benchmark)	0.72	0.71
Aindo generative model	0.77	0.74

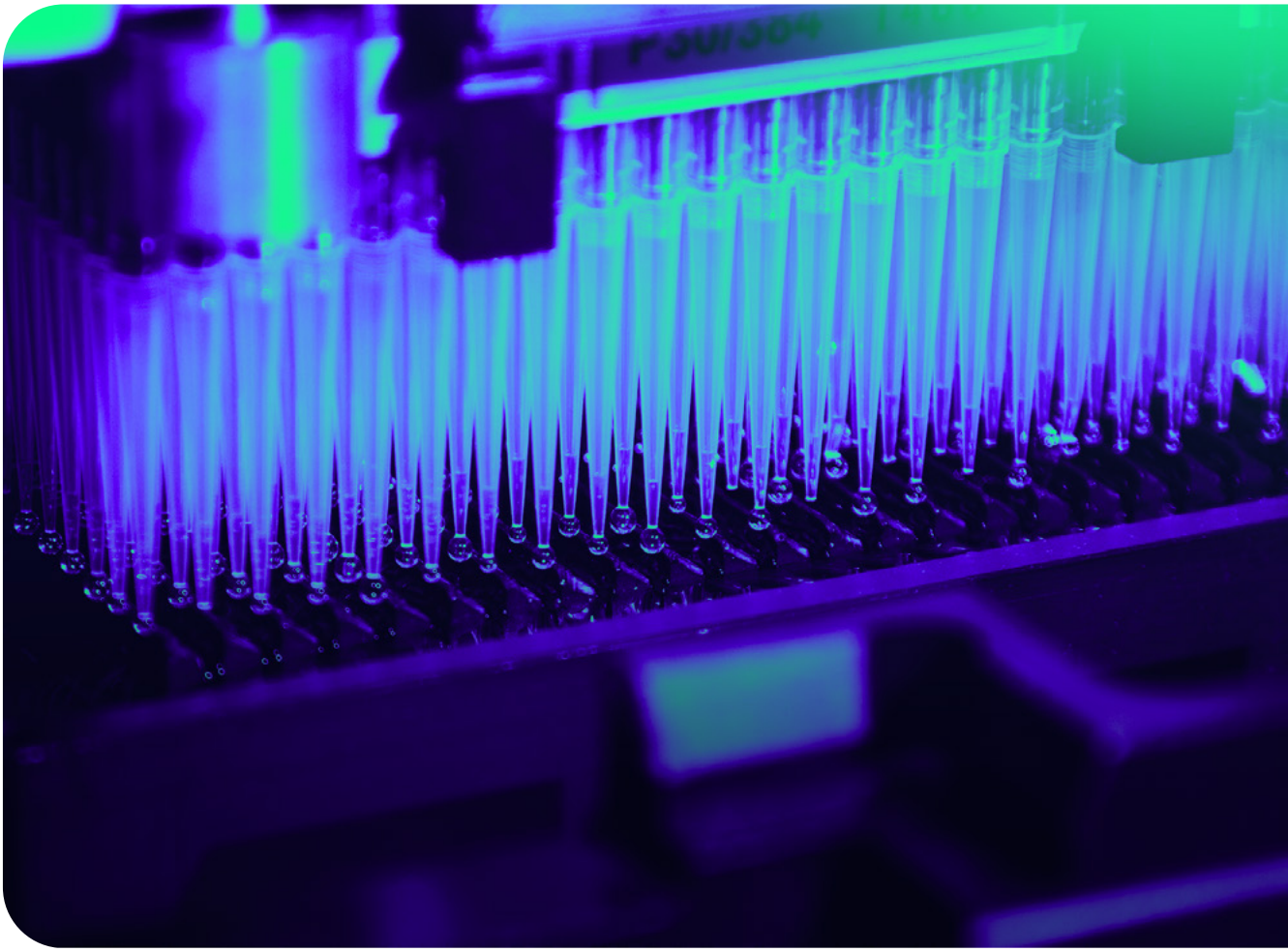
In practical terms:

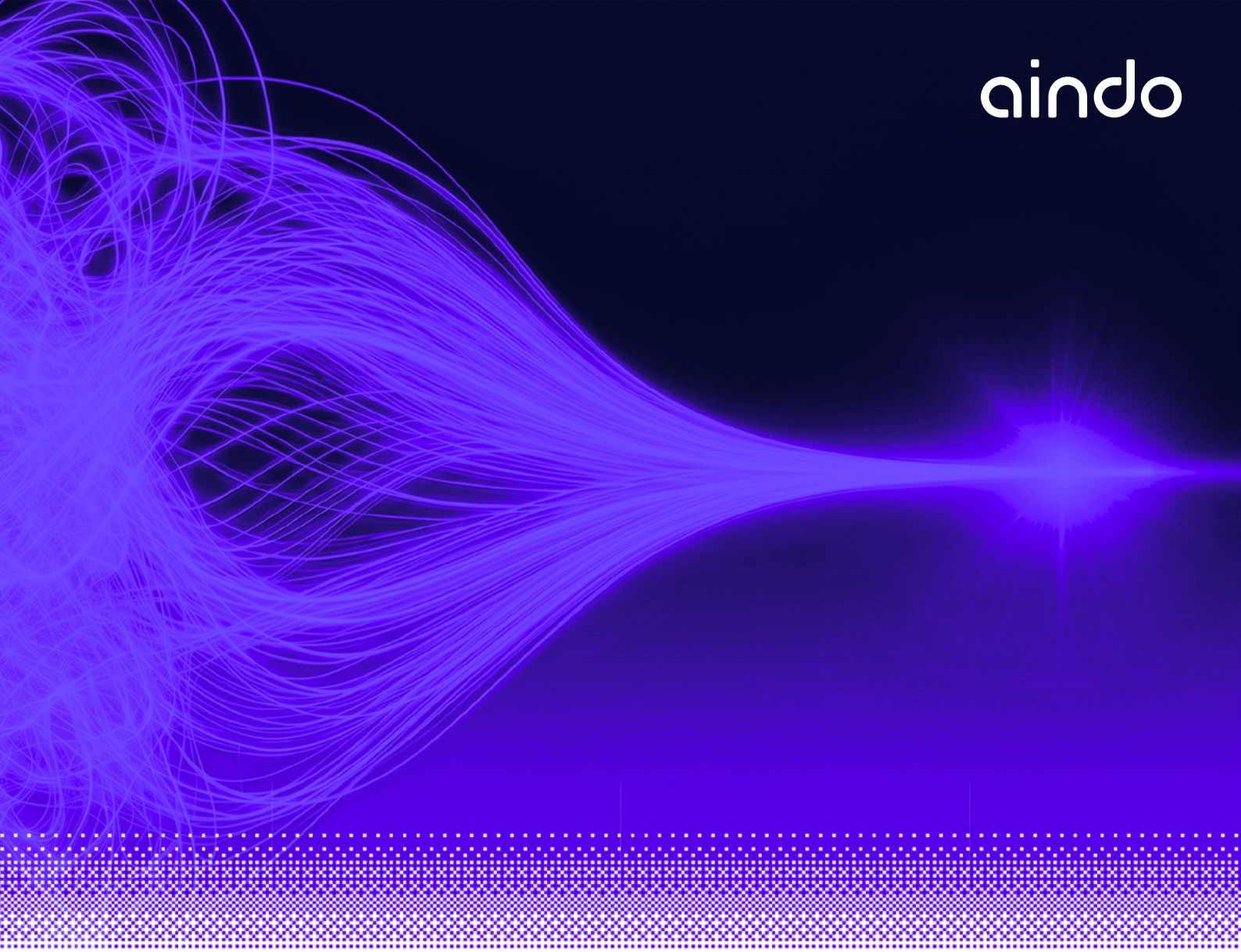
In practical terms: given one patient who reached their target and one who didn't, the Aindo model correctly identifies which is which roughly three out of every four times — a consistent advantage over a strong, modern machine learning benchmark across all cross-validation folds, and a very large one over an uninformed guess.

Why It Matters

Because the model reconstructs follow-up points that are missing or not yet collected — rather than waiting for them — it can also be applied prospectively: producing a statistically grounded estimate of what future observations are likely to show, before the corresponding real-world data has been collected in the field. For clinical research, that translates into larger effective sample sizes, stronger statistical power, and the ability to move evidence generation forward without the usual multi-year wait for a full observation window to close.

Data fidelity was assessed using the Jensen-Shannon divergence, a standard statistical measure of how different two distributions are





Thank you!

Contacts:

Alessandro Di Pietro
Head of Business Development
sales@aindo.com