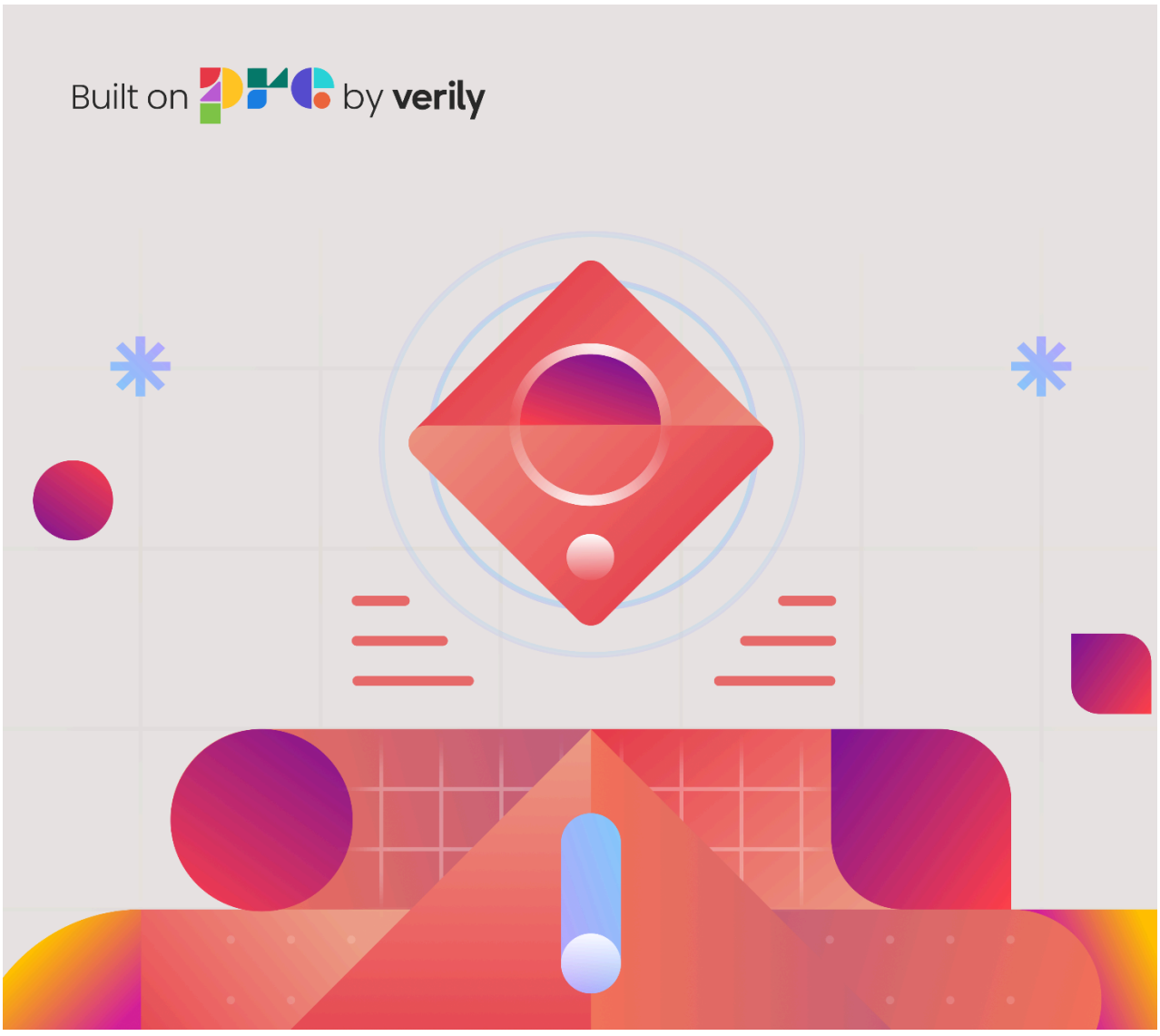
by verily

Accelerating Clinical trial recruitment with AI-driven data abstraction



Executive summary: Accelerating clinical research

Patient recruitment remains one of the most challenging bottlenecks in clinical research. While health systems sit on vast repositories of patient data, finding the right patient for the right trial at the right time is a persistent challenge.

A large Academic Medical Center (AMC) in the U.S. and Verily Health partnered to identify a scalable path to solving this challenge. As part of a pilot project, the partners assessed the value of deploying Large Language Models (LLMs) to successfully extract key variables from unstructured clinical notes across complex treatment areas, including Neurology, Cardiovascular, and Oncology.

The study evaluated how an LLM-based approach working with both structured and unstructured data could be used at-scale to identify patients eligible for clinical trials using their Electronic Health Records (EHR).

The results[1]:

- **Extraction performance:** LLM-based extraction captured 99% of target clinical values from provider notes with 94% precision (97% F1-score); 99% Recall for extracting clinical values from unstructured provider notes
- **Inference performance:** Demonstrated high precision (99%) in inferring complex eligibility criteria from provider notes when explicit values were absent; 88% Recall in identifying eligible patients through implied clinical context
- **Increased efficiency:** Deploying LLMs for this specific use case is estimated to reduce manual chart review time by 65% based on a Verily analysis

To embed this level of AI natively at an enterprise level into the patient recruitment process, the same ecosystem this pilot relied on is required: clinical data refining built with the health system's own physicians, structured human review of model output, and a secure, governed deployment environment. For organizations that want to automate and accelerate their health data workflows, Verily Pre provides these building blocks for customers to turn uncharacterized patient data into trial-ready cohorts for recruitment.

[1] Data on file

The unstructured data challenge in trial recruitment

Patient recruitment remains the well-known, biggest challenge in clinical trial operations. Up to 80% of all clinical trials fail to meet their initial enrollment targets and timelines, driving budget overruns, and delaying therapeutic advances.[2] For health systems, missed recruitment targets also mean missed revenue opportunities and, most importantly, delayed access to cutting-edge therapies for their patients. For Clinical Research Coordinators (CRCs) and Principal Investigators, locating eligible candidates is often a challenging problem.

Pain points for health systems

- **Hidden variables:** Crucial eligibility variables are often not stored in easily queried structured data fields. They must be mined from unstructured, narrative clinical notes within the EHR.
- **The burden of manual review:** CRCs are forced into exhaustive, time-consuming manual reviews of physician narratives, pathology reports, and progress notes. This can lead to staff burnout and limits the number of trials a site can effectively manage.
- **Limitations of legacy technology:**
    - Traditional Natural Language Processing pipelines often rely on manual, rule-based extractions. These systems degrade when faced with unseen or clinician-specific phrasing and cannot infer patient statuses when documentation is implicit or incomplete.
    - Approaches that require per-protocol model training impose a high maintenance burden on health system IT and data science teams, making it impossible to scale across large, diverse trial portfolios.

Solving this data challenge requires the deployment of specialized, AI-based data curation engines that can mitigate the manual review burden in an effective way.

[2] Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online Patient Recruitment in Clinical Trials: Systematic Review and Meta-Analysis. J Med Internet Res. 2020 Nov 4;22(11):e22179. doi: 10.2196/22179. PMID: 33146627; PMCID: PMC7673977.



Study overview: Extracting eligibility criteria across disease areas

The pilot study was designed to mirror research coordinator work: extracting explicit trial criteria as well as inferring statuses from context. Three experienced clinical research professionals with 6-9 years experience served as the reference standard–comparable to the coordinators this technology is meant to support.

Algorithm selection

The Verily data science team strategically decided to evaluate a zero-shot, prompt-only LLM approach for this study. This approach was chosen for trial recruitment because it eliminates the need for institution-specific, manually labeled training datasets and costly local model fine-tuning. It enables health systems to rapidly deploy and iterate on models across trials with reduced ramp-up time and lower maintenance burden.

Filtering process

The study started out with a large volume of 11,946,626 EHR notes from 20,395 patients collected at a large AMC between June 2001 and December 2024. The final study sample of 5,200 narrative progress notes was then created through a multi-step filtering process developed in partnership with clinicians at the AMC. To inform the filtering process the team reviewed note counts available in the data and clinicians highlighted what was most relevant.

The filtering process steps included

1. **Cohort creation:** Defined four disease-specific cohorts (cardiology, oncology, stroke, and hemoglobin) based on a curated set of ICD codes. Patients were required to have 3 separate, relevant ICD-10 codes documented to increase confidence that they had a target condition rather

than a single spurious or rule-out diagnosis.

2. **Additional constraints:** Leveraged a combination of data science work and clinical expertise to restrict the analysis further filtering for specific timeframes, note lengths and department association.
3. **Dataset distribution:** Split the resulting notes at the patient level into training (15%), validation (15%), and testing (70%) sets to prevent information leakage (train-test leakage).
4. **Inference evaluation only:** Removed notes explicitly mentioning ECOG and NYHA and, based on clinician recommendations, removed review of systems (ROS) notes, which sometimes appeared despite the "progress notes" filter.

Overall, the process of data filtering and preparation based on clinical knowledge was deemed important to the success of the AI-based data extraction that followed. The final analysis dataset represented **2,398 unique patients** and defined four complex, disease-specific cohorts based on ICD codes: cardiology, oncology, stroke, and hemoglobin.

Large language model configuration

The study evaluated GPT-5 and two other models using a zero-shot, prompt-only LLM configuration. All models were queried via API and their outputs were constrained to a predefined JSON schema. The prompt instructed the model to evaluate inclusion criteria directly from the clinical note text, output a boolean decision, and extract the single most recent evidence data point alongside its date, value, unit, a direct quote, and a written rationale.

Core model objectives:

1. **Extraction:** Identify explicit evidence for common trial criteria across multiple therapeutic areas (Ejection fraction, NYHA class, breast cancer stage, ECOG score, stroke documentation, and hemoglobin).

2. **Inference:** Infer functional statuses (NYHA class and ECOG performance status) when explicit documentation was missing, using a separate 630-note implicit-status corpus.

Post-processing and statistical analysis

Post-processing handled unit conversions and date arithmetic outside the LLM. Processing was conducted in a secure Google Cloud Platform environment, utilizing audit logs and a set API configuration for reproducibility.

Performance was rigorously benchmarked at the evidence, note, and patient levels against **human reference annotations**.

Human annotators were three clinical research professionals with 6-9 years of experience, comparable to real-world clinical research coordinators.

Results: Strong performance of AI-based approach

The deployed LLM achieved remarkably high accuracy for extracting trial-relevant evidence across multiple complex disease domains, proving that AI can accurately mirror–and in some cases exceed–human clinical abstraction.

Key performance metrics[3]:

Value extraction

- **Value extraction accuracy:** The model correctly identified 99% of relevant patients with 94% precision and an overall F1 score of 97%.

Inference

- **Nuanced clinical inference:** When inferring implicit NYHA or ECOG statuses based on ambiguous exertional symptoms, the LLM demonstrated substantial agreement with expert clinician judgments. When note-level decisions were aggregated per patient, the model identified 88% of relevant patients with a precision of 99% (F1 93%).
- **Complementing human benchmarks:** Manual review of errors revealed a surprising insight. For several reviewed "errors," the LLM's output was actually correct and defensible; the discrepancies were due to human annotators missing data. For example, for the NYHA functional class, instances where the model outperformed annotators emerged from capturing subtleties like “Some dyspnea with exertion.”

Table: Comparing results for clinician-reviewed errors (Objective 2 - Inference evaluation)

			Who was correct after clinician review	
	Initial Error Class	**Total**	**Annotator**	**Model**

[3] Data on file

ECOG	False negative	6	4 (66.7%)	2 (33.3%)
	False positive	10	4 (40.0%)	6 (60.0%)
NYHA Class	False negative	12	10 (83.3%)	2 (16.7%)
	False positive	3	2 (66.7%)	1 (33.3%)

Implementation considerations

The findings demonstrate that initial trial eligibility screening can be effectively automated without large labeled datasets or costly local model training. It can be deployed rapidly at health systems and research sites that lack in-house NLP expertise or specialized GPU infrastructure. Embedding this system into pre-screening workflows substantially reduces the manual hours coordinators spend reviewing charts, mitigates staff burnout, and expands the recruitment pool far beyond the obvious, structured-data candidates.

However its deployment at scale requires more extensive tooling and expertise:

- **Human experts annotating** selected model output is key to obtain consistently high performance
- **High quality monitoring and cross-functional error analysis** is needed to identify and address failure modes.
- **Protocol digitization solutions** help organizations transform PDF protocols into a digital format further accelerating the study startup process
- **Clinical informatics and data science expertise** on using structured data to filter to the right patient population is essential to make inference costs manageable

Verily's building blocks for trial recruitment

Verily provides the **subject matter expertise and technology** to support a trial recruitment use case at scale, turning uncharacterized patient data into actionable enrollment information and trial-ready cohorts.

Verily's AI-native platform for precision health, **Verily Pre,** processes diverse, multimodal data–structured and unstructured–and enables advanced analytics, workflow orchestration, and AI deployment for both care and research use cases.

- **Protocol digitization** utilizes AI to efficiently ingest and process PDF protocols, turning

pages of criteria, logic, assessments and data into a digital master study translation.

- **Refinery** operates as a high-velocity data curation engine, automating the transition from unstructured, siloed healthcare data to standardized longitudinal records and extracting high-value "invisible" variables that are often time-consuming to find or missed entirely.
- **Workbench** is an enterprise-grade Trusted Research Environment. It empowers researchers to share and co-analyze multimodal datasets, facilitating the use of powerful tools and infrastructure to turn healthcare data into insights, without compromising safety or governance.

Learn more about Refinery at
verily.com/platform
or scan the QR code to meet with an expert

About Verily

Verily is a data platform and technology company purpose-built to power AI for precision health. Verily offers AI-enabled solutions that transform disparate health data into insights and actions that accelerate research and improve care for individuals and communities. Verily is uniquely positioned at the intersection of technology, data science, and healthcare to create tools that accelerate evidence generation, enable more personalized care, and help manage disease at a population level. For more information about Verily please visit: verily.com.