Training and Validation of CARAai[™]: A Multi-LLM Platform and Data Model to Address Oncology-Specific **Challenges to Clinical Data Extraction**

Background

- In oncology, performance status, tumor characteristics, biomarkers, treatments, and tumor progression or response allow for analysis of outcomes and effectiveness.
- These concepts are derived primarily from the unstructured portion of patient EHR records.
- Historically, time and resource-intensive human abstraction was required to extract these clinical entities, resulting in extended time to insights and limited sample sizes.
- Natural language processing (NLP) has limitations in the oncology setting due to vagueness in terminology that requires understanding of the adjacent context.
- For example, "MET" could refer to the verb, a gene, or an abbreviation for "metastasis"
- The CARAai[™] platform utilizes both Large Language Models (LLMs) and Small Language Models (SLMs) across multiple AI modules to extract relevant information from patient notes and reports.

Model Training

The CARAai[™] platform uses multiple model types and orchestrates model interactions. Custom SLMs are pre-trained on medical concepts typical of medical notes in EHRs. Positive training data sets contain true labels marking the relevant clinical entities. A three-stage training process is used for all SLMs.

Figure 1. Three-stage training process for CARAai[™] SLMs. Iterations through the stages continue until high-precision models are achieved.



Figure 2. CARAai[™] Functional Architecture. The CARAai[™] platform has a series of modules to extract clinical information from unstructured EHR data. The follow of the modules occurs in the following sequence.



OCR: Document Optical Character Recognition; S-LLM: Smaller Large Language Models; LLM: Large Language Models

Jennifer R. Rider, ScD, MPH¹; Vivek Vaidya, BS²; Kuldeep Jiwani, BS²; Jeff Elton, PhD¹; Louis Culot, MLA¹; Pyeush Gurha, BS¹ ¹ConcertAl, Cambridge, MA; ²ConcertAl, Bengaluru, India

ConcertAl

Validation Methods Performance Metrics Precision = TP / (TP+FP)Recall = TP / (TP + FN)F1 = Harmonic mean of precision and recall

- Initial validation was conducted at the record level using de-identified patient records randomly sampled from 50,000 patients across 13 solid tumor types.
- De-identified patient-level validation sample size: 147 patients randomly selected from the Patient360[™] NSCLC dataset, which includes expert human abstracted clinical information.
- Patient-level validation approach:
 - Execute models to identify all unique predictions of each clinical element identified through provider notes for each patient
 - Classify each prediction as true positive (TP), false positive (FP), or false negative (FN)
 - TP = CARAai[™] prediction matches human abstraction
 - FP = CARAai[™] prediction not identified by human abstraction following verification by review of subject matter expert
 - FN = CARAai[™] prediction not identified for an element available by human abstraction
- Support = TP identified by initial human clinical curation plus the conversion of FP to TP after human verification

Results

Table 1. Performance metrics of CARAai[™] at a patient level compared to human abstraction among 147 randomly selected patients included in the Patient360[™] NSCLC dataset.

Clinical element	Occurrences identified through expert clinical curation	Occurrences identified through CARAai™ platform	TP	FP	FN	Support	Precision	Recall	F1
		Tumor char	acteristic	S					
Primary cancer	147	126	126	0	21	147	1	0.86	0.92
AJCC stage	121	313	258	56	7	265	0.82	0.97	0.89
AJCC T stage	84	162	155	8	6	161	0.95	0.96	0.95
AJCC N stage	72	158	155	3	3	158	0.98	0.98	0.98
AJCC M stage	77	139	133	6	3	136	0.96	0.98	0.97
Tumor grade	63	225	117	102	2	119	0.52	0.98	0.66
Histology	162	871	632	129	19	651	0.83	0.97	0.92
		Clinical cha	racteristic	S					
Alcohol assessment	105	144	133	11	20	153	0.92	0.87	0.89
Smoking status	139	147	128	19	38	166	0.87	0.77	0.82
Comorbidities	99	229	208	21	39	247	0.91	0.84	0.87
Biomarker name	362	448	421	27	18	439	0.94	0.96	0.95
		Cancer treatmen	t and out	comes					
Medication	134	551	482	69	0	482	0.87	1.00	0.93
Treatment response	260	704	414	143	39	453	0.74	0.91	0.82

Key Takeaway

The CARAai™ LLM suite achieved high patient-level precision and recall relative to expert human clinical abstraction for key oncology data elements, achieving non-inferiority to human methods. The CARAai™ LLM models will facilitate improved statistical power and timeliness for HEOR and epidemiologic studies on outcomes and safety.

RWD116

ConcertA