Extraction of Pulmonary Function Test Results from Unstructured Clinical Notes using a Retrieval-Augmented Generation Approach on a Major Cloud Platform

Background

- Disease severity refers to the extent, intensity, or seriousness of a disease and is often measured by clinical signs, symptoms, laboratory values, and/or other outcomes.
- Pharmacologic and non-pharmacologic management or disease is often highly dependent on severity.
- Severity measures are not always available in structured electronic health record (EHR) data; however, they may be embedded in clinical notes.
- Natural language processing (NLP) methods have been applied in recent years to extract severity measures from clinical notes.
- Large language models (LLMs) represent a newer NLP methodology for extracting severity scores from clinical notes; however, their accuracy and hallucination rate fo this use case is unclear.

Objective

 To measure the accuracy, hallucination rate, and latency with which large language models (LLMs) can extract the forced expiratory volume in one second to functional vital capacity ratio (FEV1/FVC) from relevant electronic health record free-text clinical notes.

Methods

- 50 note excerpts containing the string "FEV1/FVC" from the OMNY Health real-world data platform were randomly sampled and classified as simple (S; one FEV1/FVC score present), no value (NV; no scores present), or complex (C; multiple scores present).
- Two LLMs [Gemini-1.5-Flash (Flash) and Gemini-1.5-Pro(Pro)] available in Google BigQuery ML Studio were applied to query the notes using the following prompt: "Extract the actual FEV1/FVC ratio from the following text. Text: <NOTE>."
- Maximum output tokens, temperature, and top-P parameters were 4, 0, and 0, respectively.
- Results for the S/NV categories were annotated for accuracy and hallucinations by a team of domain experts following a standardized protocol. C category results were qualitatively evaluated, and latency for each model was measured.

	Results
	 The S, NV, and C categories comprised 30 (60%), (18%), and 11 (22%) excerpts, respectively.
ſ	 Example note excerpts and results from the Flash and Pro LLMs with accompanying annotations ar
DŢ	presented in Table 1.
d	Many notes provided the component measurements (FEV1 and FVC) in addition to th ratio.
	–Some notes provided multiple values for the FEV1/FVC ratio.
	 Relative accuracy and latency of the Flash and Pr models for FEV1/FVC extraction from clinical not is presented in Figure 1:
l or	–For the S category, Flash and Pro models were 90.0% and 73.3% accurate, respectively.
	 For the NV category, no hallucinations were observed.
	-For the C category, the Flash model reported or of the FEV1/FVC values, while the Pro model
	recognized the presence of multiple values.
	milliseconds, respectively.

Conclusions

- LLMs can extract severity scores with 90% accuracy from simple note excerpts automatedly scale.
- The Flash model was more accurate with simple excerpts, while the Pro model was more verbose (limiting its utility for when the max output tokens low) yet more cognizant of multiple FEV1/FVC values.
- Potential future directions include exploring additional metrics beyond accuracy and latency, such as robustness across note complexity levels and model interpretability while accounting for co as a factor, as well as model fine-tuning.

Vikas Kumar | Yan Wang | Lawrence Rasouliyan OMNY Health, Atlanta, GA

	Text	Excerpt Type	Flash Result	Flash Correct?	Pro Result
Pulmonary fur 2.78 (98 % pre predicted), R\ predicted)	nction tests: FEV1: 2.14 (90 % predicted), F edicted), FEV1/FVC: 77 , TLC: 4.79 (110 % / 1.91 (126 % predicted), DLCO: 20 (68 %	VC: Simple	77	Y	77
PFT's Singles FEF 25-75: 80	pirometry FVC: 88 FEV1: 88 FEV1/FVC: 88 Interpretation: Normal PFT	Simple	88	Y	"The provided text savs"
PULMONARY Reviewed <da restrictive pat FEV1 35% pre predicted from 37% predicted predicted and postbroncho predicted and 54% predicted predicted, but Airway resista</da 	FUNCTION TEST: [] Ordered [] Unavailable ATE> Â The patient's spirometry shows a tern with a significant bronchodilator respo dicted with 54% bronchodilator change to 5 n 1.01 liters to 1.56 liters. Â FVC 1.37 liters d with 54% bronchodilator change to 57% 2.12 liters. Â FEV1/FVC ratio pre and dilator is 74. Â Total lung capacity 105% RV is 183% predicted and slow vital capacit d and diffusion capacity uncorrected is 59% corrected for alveolar volume is 87% predicted ance is significantly increased.	[x] nse, 54% , Complex ity	74		74
PFT Values otherwise, on displayed. 88% 84% FE 77% TLC 109	Some values may be hidden. Unless noted ly the newest values recorded on each date Old Values PFT Results <date> <date> F /1 77% 68% FEV1/FVC 69.32% 64.06% DI</date></date>	are VC Complex LCO	64	N	"Two FEV1"
PFT interpreta Mild obstructi	tion: Maneuver: valid and meets ATS guidel on with decreased FEV1 and FEV1/FVC	ines No Value	"The provided text does"	Y	"The provided text does'
PFT interpreta Mild obstruction Figure 1. Re Google Clo 100	tion: Maneuver: valid and meets ATS guidel on with decreased FEV1 and FEV1/FVC elative Accuracy (for "Simple" Cate ud Platform Models for FEV1/FVC	ines No Value egory) and La Extraction	"The provided text does"	Y	"The provided text does" tegories
PFT interpreta Mild obstruction Figure 1. Re Google Clo 100	tion: Maneuver: valid and meets ATS guidel on with decreased FEV1 and FEV1/FVC elative Accuracy (for "Simple" Cate ud Platform Models for FEV1/FVC	ines No Value	"The provided text does"	Y	"The provided text does' tegories
PFT interpreta Mild obstructi Figure 1. R Google Clo 100 80 80 80	tion: Maneuver: valid and meets ATS guidel on with decreased FEV1 and FEV1/FVC elative Accuracy (for "Simple" Cate ud Platform Models for FEV1/FVC	ines No Value egory) and La Extraction	"The provided text does"	Y	"The provided text does" tegories
PFT interpreta Mild obstructi Figure 1. R Google Clo 100 80 80 60 0000 40	tion: Maneuver: valid and meets ATS guidel on with decreased FEV1 and FEV1/FVC elative Accuracy (for "Simple" Cate ud Platform Models for FEV1/FVC	ines No Value egory) and La Extraction	"The provided text does"	Y	"The provided text does"
PFT interpreta Mild obstruction Figure 1. R Google Cloo 100 80 80 80 20 20	tion: Maneuver: valid and meets ATS guidel on with decreased FEV1 and FEV1/FVC elative Accuracy (for "Simple" Cate ud Platform Models for FEV1/FVC	ines No Value egory) and La Extraction	"The provided text does"	Y for All Ca	"The provided text does tegories

