Supplemental Online Content

Cunningham JW, Singh P, Reeder C, et al. Natural language processing for adjudication of heart failure in a multi-center clinical trial: a secondary analysis of a randomized clinical trial. *JAMA Cardiol*. Published online November 11, 2023. doi:10.1001/jamacardio.2023.4859

This supplemental material has been provided by the authors to give readers additional information about their work.

eMethods

Optical Character Recognition

Adjudication dossiers were converted from Portable Document Format (PDF) to TIFF images files using Ghostscript Seamless and from TIFF to text using Tesseract optical character recognition (Supplemental Methods). Some PDFs had character data (able to be highlighted or copied) while others were images. Our process converted PDFs with character data to image files (TIFF format). Most personal health information such as patient name and date of birth had already been redacted from the dossiers by site research staff as part of the INVESTED CEC process. The formatting of tables was not consistently retained. Headers and footers were included in the plain text output. Some dossiers contained images, most commonly 12-lead electrocardiograms. The images themselves were not converted, but text on the images (such as the automated reading at the top of an electrocardiogram) was converted. We were careful to eliminate the cover sheet, the only piece of substantive information added by the trial process. Some dossiers from early in the trial period had multi-page cover sheets, which were identified and eliminated. In general, the medical records did not contain additional INVESTED specific markings or information, though occasionally the study ID number was handwritten on some pages by research staff; these annotations were not removed. On side-by-side manual review of 30 dossier PDFs and the plain text output of the optical character recognition pipeline, the plain text matched the PDF text with very high accuracy in all cases.

Development and Validation of the C3PO NLP Model for HF Hospitalization

The development and validation of the C3PO NLP model is presented in our previous publication (Cunningham JW et al, JACC Heart Failure, 2023). Briefly, cardiologists adjudicated 1934 discharge summaries from hospitalizations with ICD codes for HF in C3PO, an electronic health record cohort of patients receiving longitudinal primary care at Mass General Brigham. We focused on discharge summaries because they were available consistently electronically over time and provide a parsimonious summary of hospital course. We trained multiple transformer-based NLP models based on different pre-training architectures in a training set (n=1268). The development set included 214 hospitalizations with ICD codes for heart failure selected from the C3PO cohort. Each architecture was trained for 15 epochs in the identical training set, and cached to maximize average precision on the test set. Batch size of all BERT based models was 32, while batch size for Longformer-based models was 4. All other hyperparameters were held consistent across training jobs.

In a development set (n=214), the best model was based on Clinical Longformer pre-training (eTable 1), trained for 15 epochs. The NLP model produces a continuous score reflecting likelihood of HF; the best threshold for binary adjudication of HF hospitalization was defined in the development set as 0.958. In a held-out internal validation set, adding the C3PO NLP HF model to ICD codes improved adjudication accuracy compared to clinician review.

eTable 1. Development Test Set Performance of Each Pre-Trained NLP Model During Original Development of the C3PO NLP Model

Model Architecture	Average Precision	Area Under ROC
BERTBASE	0.756	0.815
Bio+DischargeSummaryBERT	0.784	0.846
PubMedBERT	0.793	0.856
SapBERT	0.768	0.849
LongformerBASE	0.786	0.854
Clinical Longformer	0.884	0.933

Development of the Fine-Tuned and *de novo* Retrained INVESTED NLP Models

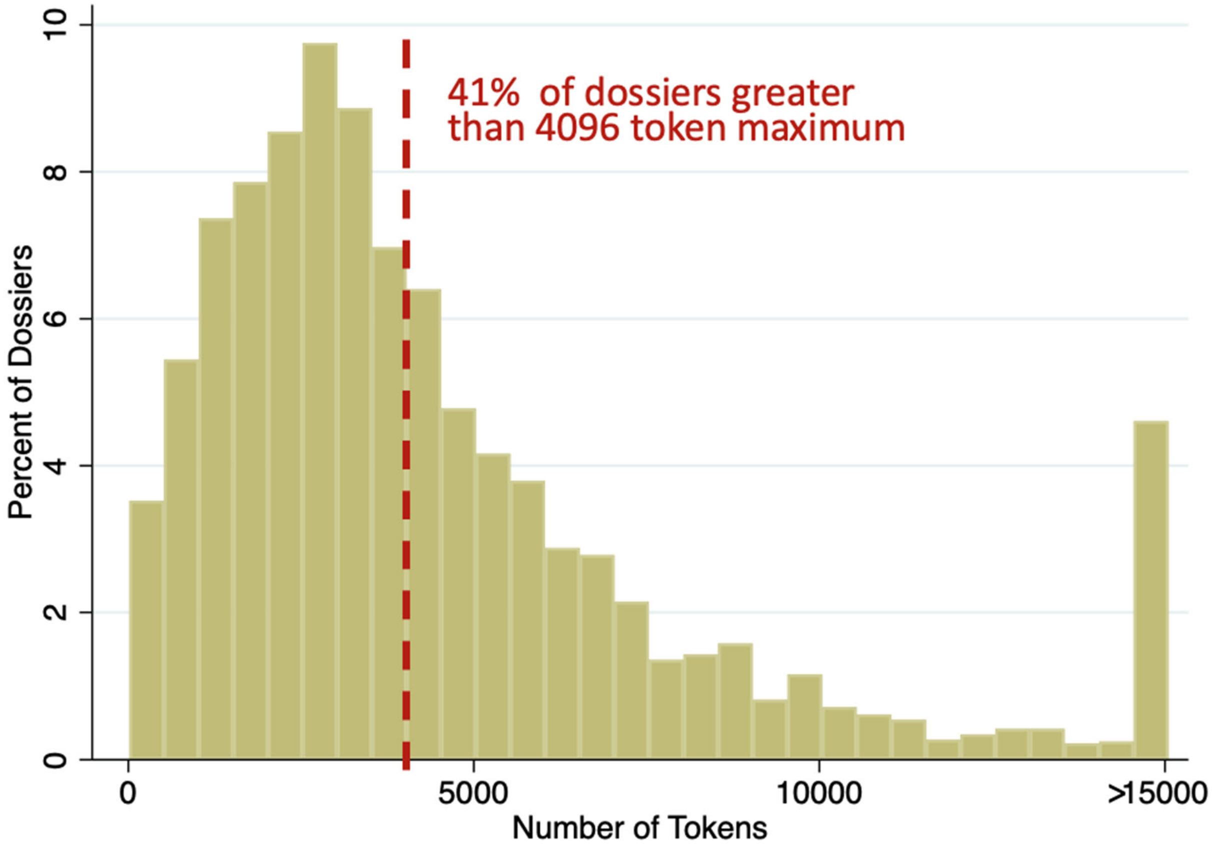
The INVESTED NLP models were trained in the development subset of INVESTED based on the Clinical Longformer pre-trained architecture for up to 5 epochs. The best model was chosen based on average precision in a 200-400 hospitalization subset of the development set which was not used for initial model training. Models were evaluated every 500 steps. The Adam Optimizer was used. The learning rate was 10^{-5}.

eTable 2. Rate of NLP Heart Failure by True CEC Adjudication

CEC Adjudication	Total Hospitalizations	NLP HF Hospitalizations	% NLP HF
Heart Failure	1,074	1,009	94%
Cardiopulmonary Non-Specific	300	199	66%
Non-HF Cardiovascular Causes	1375	297	22%
Pulmonary	290	31	11%
Non-Cardiopulmonary	1,305	147	11%
Unknown	16	2	13%

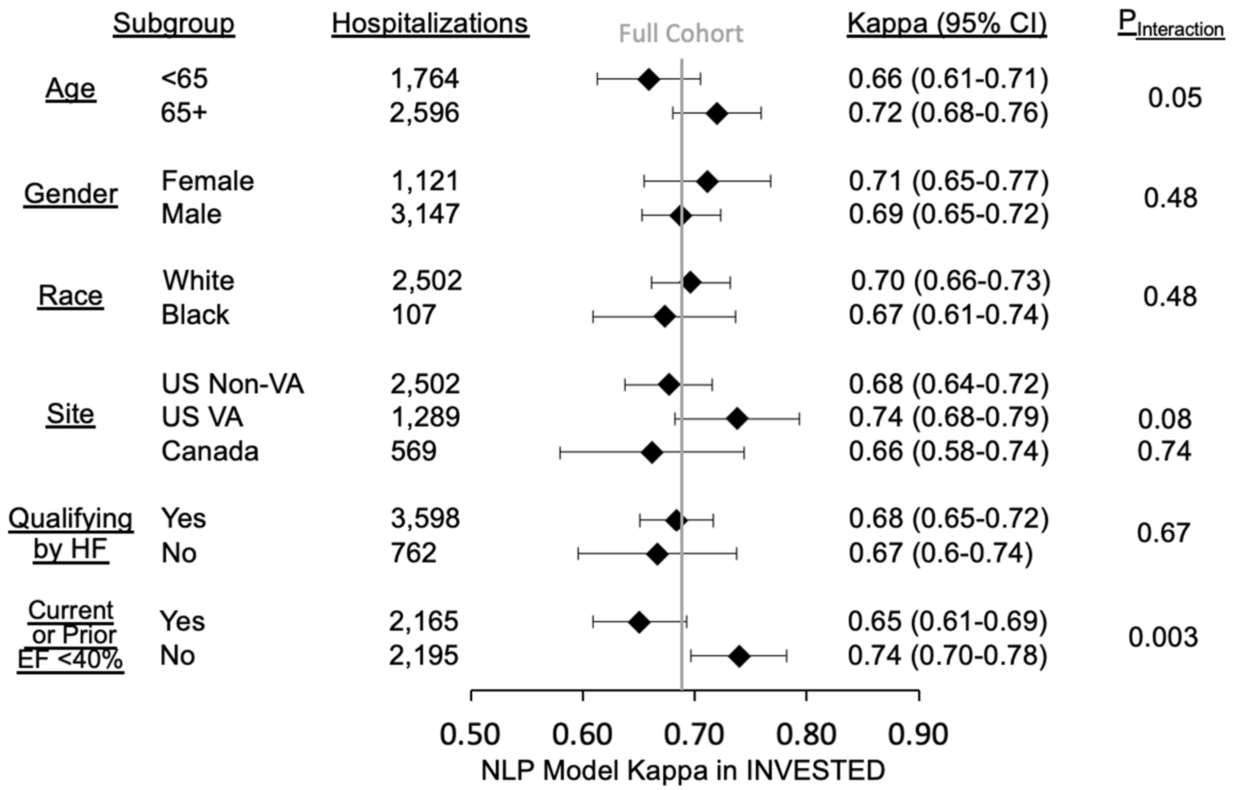
HF, heart failure; NLP, natural language processing.

eFigure 1: Histogram of Token Length of Medical Record Dossiers



Total sample size for this histogram is n=4060, the total hospitalizations in the primary analysis

eFigure 2: Agreement Between NLP and Human CEC Heart Failure Adjudications in Key Subgroups of Patients



Subgroup		Hospitalizations	Kappa (95% CI)	$P_{Interaction}$
Age	<65	1,764	0.66 (0.61-0.71)	0.05
	65+	2,596	0.72 (0.68-0.76)	
Gender	Female	1,121	0.71 (0.65-0.77)	0.48
	Male	3,147	0.69 (0.65-0.72)	
Race	White	2,502	0.70 (0.66-0.73)	0.48
	Black	107	0.67 (0.61-0.74)	
Site	US Non-VA	2,502	0.68 (0.64-0.72)	
	US VA	1,289	0.74 (0.68-0.79)	0.08
	Canada	569	0.66 (0.58-0.74)	0.74
Qualifying by HF	Yes	3,598	0.68 (0.65-0.72)	0.67
	No	762	0.67 (0.6-0.74)	
Current or Prior EF <40%	Yes	2,165	0.65 (0.61-0.69)	0.003
	No	2,195	0.74 (0.70-0.78)	

Interaction p-value evaluates the null hypothesis that the kappa statistic does not differ between subgroups. CI, confidence interval. EF, ejection fraction; HFH, heart failure hospitalization; VA, Veterans Administration.