APPENDIX 1

(details of natural language processing)

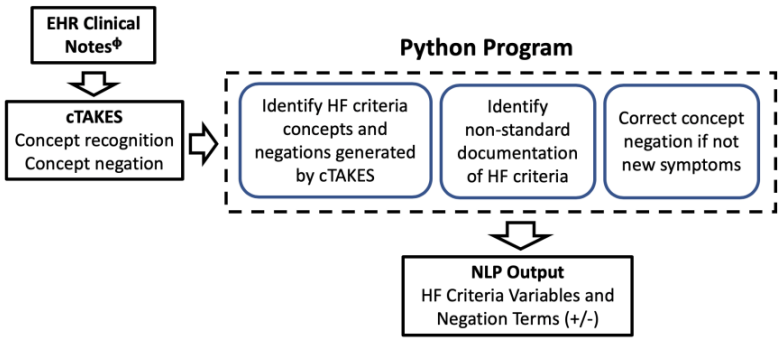
Natural Language Processing (NLP)

cTAKES NLP Program

The clinical Text Analysis and Knowledge Extraction Tool (cTAKES) is open source natural language processing (NLP) package that is specifically designed to analyze clinical free-text notes and reports; cTAKES uses specific modules for clinical concept coding, as well as negation status of identified clinical concepts. Concept coding based on standardized electronic dictionaries (UMLS Metathesaurus) are used by cTAKES to identify clinical terms of interest, such as 'dyspnea', and assign to them the appropriate concept unique identifier (CUI), such as 'C0013404', that can be identified by a computer algorithm from cTAKES output files. The cTAKES program also assigns a negation status to each concept it identifies in the electronic text. For example, the term 'no dyspnea' is processed by cTAKES by first assigning the concept unique identifier (CUI), 'C0013404', to the 'dyspnea' term and then assigns a negation flag to the concept.

The project required an additional layer of negation so atypical negations not captured by cTAKES were identified. Therefore, we developed a program written in Python that processes the cTAKES output (and raw text) and negates relevant concepts not captured by cTAKES (Figure 1).

Figure 1. ARIC Natural Language Processing Pipeline (NLP)



ϕ emergency department notes, hospital admission notes, and discharge summaries
cTAKES: clinical Text Analysis Knowledge Extraction System
Negation: determination of whether a HF criteria is negated (e.g., patient has edema vs. patient has no edema)

For example, the 'dyspnea' concept in the text 'the patient has chronic dyspnea' is negated by the Python program because the symptoms are not new onset or worsening. In this example, the term 'chronic' is recognized by the Python program as negating 'dyspnea'.

Modification of cTAKES Negation of Framingham HF Criteria Variables

Table 1 shows the CUIs and negation terms (added to cTAKES negation) used to identify relevant Framingham criteria HF variables. For example, dyspnea on exertion, three CUIs identify this clinical symptom; 'C0013404', 'C0231835', and 'C1145670'. These CUIs represent 'shortness of breath', 'tachypnea', and 'respiratory distress'; respectively. These terms are all synonymous for 'dyspnea on exertion'. Their respective negation terms are shown in the last column. Again, these are negations that augment cTAKES negation. The program allows the user to specify the number of characters before, and after, the relevant Framingham HF criteria term to search for the relevant negation term.

Table 1. List of relevant CUI codes for Framingham HF Criteria Variables and Negation Terms

Description	CUI	CUI Description	Negation Terms (added to cTAKES negation)
Dyspnea on exertion	C0013404	Dyspnea, shortness of breath	'negative', 'neg', 'denies', 'denied', 'needed', 'stable', 'prn', 'chronic', 'without', 'or', 'if', 'mild', 'history', 'hx', 'allergy'
	C0231835	tachypnea	'neg', 'denies', 'denied'
	C1145670	Respiratory failure	'neg'
Lower extremity edema	C0013604	Edema	'negative', 'neg', 'denies', 'denied', 'history', 'hx', 'h/o', 'chronic', 'stable', 'mild', 'trace', 'if', 'or', 'infiltration', 'related', 'cxr', 'lobe', 'alveolar', 'facial', 'for', 'pulmonary', 'interstitial', 'mass', 'pulm', 'hemorrhage', 'represent', 'mesenteric', 'residual', 'arm', 'vasogenic', 'gallbladder', 'w/o'
	C0239340	Lower extremity edema	'negative', 'neg', 'denies', 'denied', 'trace', 'pulmonary', 'resolved', 'chronic', 'stable', 'needed', 'prn', 'without', 'history', 'hx', 'h/o'
	C0581394	Lower extremity swelling	'negative', 'neg', 'denies', 'denied', 'trace', 'pulmonary'
	C0235439	Swollen ankles	'negative', 'neg', 'denies', 'denied'
	C0151603	Anasarca	'negative', 'neg'
Jugular venous distention (JVD)	C0425687	jugular venous distention	'negative', 'neg'
Hepatojugular reflux	C0239949	hepatojugular reflux	'neg'
Hepatomegaly	C0019209	hepatomegaly	'neg'
Paroxysmal nocturnal dyspnea	C1956415	paroxysmal nocturnal dyspnea	'negative', 'neg', 'denies', 'denied'
	C0344357	nocturnal dyspnea	'negative', 'neg', 'denies', 'denied'
Orthopnea	C0085619	orthopnea	'negative', 'neg', 'denies', 'denied'
Pulmonary basilar rales	C0034642	rales	'negative', 'neg', 'denies', 'denied'
S3 gallop	C0232278	S3 gallop	'neg'
Alveolar/Pulmonary edema on chest x-ray	C0034063	Pulmonary edema	'versus', 'vs', 'neg', 'history', 'mild', 'may', 'could', 'possible'
	C0748120	Pulmonary interstitial edema	'neg'
	C2939069	Alveolar edema	'neg'
	C4313225	Interstitial infiltrates	'neg'
	C2750120	Interstitial thickening	'neg'
	C3670573	Perivascular edema	'neg'
	C0746171	Lower lung interstitial markings increased	'neg'
Cardiomegaly	C0018800	cardiomegaly	'neg'
Bilateral pleural effusion	C0032227	pleural effusion	'resolution', 'possible'

Handling of Non-traditional ways of Documenting Framingham HF Criteria Variables in Clinical Text

Review of the clinical notes and reports revealed several idiosyncratic ways in which clinicians document HF signs and symptoms. In these instances, cTAKES does not recognize the terms as clinical concepts and therefore does not assign CUIs. For example, 'jugular venous distention' is often abbreviated 'JVD' in clinical text. However, 'JVD' is not recognized by cTAKES as a clinical concept and not assigned a CUI. Therefore, we build a module in the Python program that used regular expression to identify non-traditional terms and assign them to the correct CUI.

Detail of this process are as follows:

1. Jugular venous distention (JVD) is sometimes documented as increased jugular venous pressure (JVP). JVP does not have a CUI and therefore regular expressions were used to identify this sign. When the term JVP is identified, the python program assigns it the JVD CUI (C0425687) and then performs negation as follows:

   Negation_terms_left_jvd and negation_terms_right_jvd applied on top of negation_terms_left/right terms to jvd/jvp specific CUI extracted manually i.e. not generated by cTAKES

   - negation_terms_left_jvd = ['no jvp', 'unable to', 'could not', 'normal jvp', 'negative', 'neg', 'no jvd', 'unable to', 'could not', 'normal jvd', 'negative', 'neg']
   - negation_terms_right_jvd = ['jvp not', 'unelevated', 'jvp normal', "jvp doesn't", 'jvp does not', 'jvd not', 'unelevated', 'jvd normal',"jvd doesn't", 'jvd does not']
2. Paroxysmal nocturnal dyspnea is sometimes documented as PND. cTAKES does not recognize PND (no CUI assigned); therefore, regular expressions were used to identify this term. When the term PND is identified, the python program assigns it the paroxysmal nocturnal dyspnea CUI and then performs negation as follows:

   Negation_terms_left_pnd and negation_terms_left_negs applies on top of negation_terms_left/right terms to pnd specific CUI extracted manually i.e. not generated by cTAKES

   - negation_terms_left_pnd = ['no pnd', 'no paroxysmal nocturnal dyspnea', 'denies pnd', 'denies paroxysmal nocturnal dyspnea', 'denied pnd', 'denied paroxysmal nocturnal dyspnea']
   - negation_terms_left_negs = ['neg']

   Additionally, if the program identifies a ‘shortness of breath’ CUI, it searches for the following terms in close proximity: 'woke', 'overnight', 'sleep', and 'awakened'. If it finds any of these terms, then ‘paroxysmal nocturnal dyspnea’ is considered to be present.