Submission type: Original article

Title: Deep natural language processing to identify symptom documentation in clinical notes for patients with heart failure undergoing cardiac resynchronization therapy

Richard E. Leiter, MD, MA[1,2,3]*

Enrico Santus, PhD[4]*

Zhijing Jin, BEng[4]

Katherine Lee, MD[5]

Miryam Yusufov, PhD[1,2]

Isabel Chien, MEng[4]

Ashwin Ramaswamy, MD[6]

Edward T. Moseley, BS[2]

Yujie Qian, BEng[4]

Deborah Schrag, MD, MPH[1,7]

Charlotta Lindvall, MD, PhD[1,2,3]

*Drs Leiter and Santus are co-first authors, contributed equally to the study, had full access to all of the data in the study, and take responsibility for the integrity of the data and accuracy of the data analysis.

Corresponding Author:
Richard E. Leiter, MD, MA
450 Brookline Ave., Jimmy Fund 8
Jimmy Fund 8
Boston, MA 02215 USA
Phone: 617-632-6464
Fax: 617-632-6180
Email: richard_leiter@dfci.harvard.edu

[1]Harvard Medical School, Boston, Massachusetts
[2]Department of Psychosocial Oncology and Palliative Care, Dana-Farber Cancer Institute, Boston, Massachusetts
[3]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts

[4]Massachusetts Institute of Technology, Boston, Massachusetts
[5]Department of Surgery, University of California San Diego Health, San Diego, California
[6]Department of Surgery, NewYork-Presbyterian Hospital/Weill Cornell Medical Center, New York, New York
[7]Division of Population Sciences, Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts

Word Count: 3500 (limit 3500)

Abstract

Word Count: 250 (limit 250)

Context: Clinicians lack reliable methods to predict which patients with congestive heart failure (CHF) will benefit from cardiac resynchronization therapy (CRT). Symptom burden may help to predict response, but this information is buried in free-text clinical notes. Natural language processing (NLP) may identify symptoms recorded in the electronic health record (EHR) and thereby enable this information to inform clinical decisions about the appropriateness of CRT.

Objectives: To develop, train, and test a deep NLP model that identifies documented symptoms in CHF patients receiving CRT.

Methods: We identified a random sample of clinical notes from a cohort of patients with CHF who later received CRT. Investigators labeled documented symptoms as *present, absent,* and *context-dependent* (pathologic depending on the clinical situation). The algorithm was trained on 80% and fine-tuned parameters on 10% of the notes. We tested the model on the remaining 10%. We compared the model's performance to investigators' annotations using accuracy, precision (positive predictive value), recall (sensitivity), and F1 score (a combined measure of precision and recall).

Results: Investigators annotated 154 notes (352,157 words) and identified 1340 present, 1300 absent, and 221 context-dependent symptoms. In the test set of 15 notes (35,467 words), the model's accuracy was 99.4% and recall was 66.8%. Precision was 77.6% and overall F1 score was 71.8. F1 scores for present (70.8) and absent (74.7) symptoms were higher than that for context-dependent symptoms (48.3).

Conclusion: A deep NLP algorithm can be trained to capture symptoms in CHF patients who received CRT with promising precision and recall.

Keywords: Artificial Intelligence; Heart Failure; Cardiac Resynchronization Therapy; Signs and Symptoms; Clinical Decision-Making

Key Message (word count 49; limit 50): This article describes the development, training, and testing of a deep natural language processing model that identifies documented symptoms in patients with congestive heart failure receiving cardiac resynchronization therapy. Results indicate that our model can be trained to capture symptoms in this patient population with promising precision and recall.

Running Title (42 characters; limit 50): Deep NLP to Identify Symptoms Prior to CRT

1.0 Introduction

Patients with advanced heart failure with reduced ejection fraction (HFrEF) have a high symptom burden (1,2) and limited life expectancy, and may benefit from palliative care involvement (3–7). Palliative care expertise can be particularly valuable in HFrEF-related medical decision-making, such as decisions cardiac resynchronization therapy (CRT), which involves implanting a biventricular pacemaker with or without an implantable cardiac defibrillator (ICD). CRT is an established therapy for patients with both severe HFrEF and dyssynchronous ventricular activation(8–11), the latter of which is associated with reduced cardiac output and increased mortality(8,9,12,13). CRT leads to increased left ventricular ejection fraction (LVEF), decreased heart failure-related hospitalizations, improved survival, and higher quality of life in some, but not all patients (14–22). Identifying the patients who are most likely to benefit and the optimal timing for CRT insertion remains challenging. Premature CRT insertion puts patients at higher risk of infection and inappropriate defibrillation and delayed insertion compromises survival(11,12,23–26). However, clinicians are often ill-equipped to make informed recommendations about CRT.

Symptom burden and its trajectories hold unrealized potential to differentiate patients who will benefit from those who are unlikely to benefit from the device. Patients with HFrEF experience a significant symptom burden (20, 21) including those directly (e.g. chest pain and dyspnea) and indirectly (e.g. nausea, anorexia, and depressed mood) related to HFrEF (27). These symptoms are common, but clinicians struggle with recognizing the most symptomatic patients. While researchers commonly use validated patient-reported outcome (PRO) tools such as the Kansas City Cardiomyopathy Questionnaire (KCCQ) and the Minnesota Living with Heart Failure Questionnaire (MLHFQ)(28,29), clinicians face multiple barriers to integrating these

tools into practice. Because structured information related to a patient's symptom burden is unavailable, there are missed opportunities to identify appropriate CRT candidates.

The electronic health record (EHR) represents one possible way forward; however, until recently, extracting symptoms from the free text of clinical notes necessitated manual chart abstraction, which is too inefficient and time-consuming to be a viable means of identifying symptoms. Newer artificial intelligence-based methods such as deep learning hold promise to improve this process. When used together, natural language processing (NLP) and deep learning algorithms can ascertain flexible and generalizable patterns from a labeled set of free-text notes and then apply those rules to different, unlabeled notes(30–33). This method is ideal in situations where the set of extraction rules is large, unknown, or both, as is the case when clinicians describe patient symptoms in clinical notes. Our objective in this study was to develop, train, and evaluate a deep NLP algorithm to identify documented symptoms from unprocessed (i.e. not edited by investigators) EHR notes in a cohort of patients with HFrEF who would subsequently undergo CRT.

2.0 Methods

2.1 Data and Algorithm

Data Source

Our primary data source was the Partners HealthCare Research Patient Data Registry(34,35). The Research Patient Data Registry includes data from 4.6 million patients collected over more than 20 years from multiple EHR systems at Partners HealthCare, a large healthcare system in Eastern Massachusetts. The database contains more than 227 million encounters, 193 million billing diagnoses, 105 million medications, 200 million procedures, 852

million lab values, and over 5 million unstructured clinical notes, including outpatient visit notes, inpatient admission and consultation notes, cardiology reports (e.g. EKGs and echocardiograms), and others. Patient-level data are available for research after peer-review of project proposals. The Partners Human Research Committee approved this study.

Study Population

Patients who underwent CRT implantation between January 2004 and December 2015 at Brigham and Women's Hospital and Massachusetts General Hospital were eligible. We identified cases using *International Classification of Diseases, Ninth Revision* (ICD-9) and Current Procedural Terminology (CPT) codes for CRT: ICD-9 00.50, ICD-9 00.51, CPT 33224, CPT 33225, or CPT 33226. We included patients who underwent initial implantation of either a CRT pacemaker (CRT-P) or CRT with ICD (CRT-D). As this dataset was also used for a parallel study examining LVEF response to CRT placement, we excluded patients who lacked baseline measurement of LVEF within 60 days of the procedure or follow-up LVEF measurement between 6 and 18 months post-procedure and were alive at 18 months, or who received a CRT within 18 months of the end of the dataset's time window.

From this population, we selected a random sample of discharge summaries from hospitalizations that were either prior to CRT-implantation, or from the admission during which the patient underwent the procedure. We chose discharge summaries as each note contains an admission History of Present Illness (HPI) and Review of Systems (ROS) that reflects the patient's symptom burden prior to undergoing CRT implantation during that hospitalization. We had initially planned to also examine the Hospital Course, but concluded that these contain a mixture of symptoms and signs and were thus less useful for initial evaluation of the algorithm.

Annotation Software

PyCCI is a local executable software program built using Python Version 3.5. For this project, we created a customized version of PyCCI to support symptom annotation. Features include color highlighting and inclusion of associated modifiers to annotate symptoms. PyCCI's use during the annotation process allows for the collection of annotation information necessary for the utilization of modern NLP algorithms, including information for benchmarking annotation speed comparisons between humans and machines.

Symptom Coding and Categorization.

Prior to coding, we convened an expert panel of palliative care clinicians and researchers, computer programmers, artificial intelligence and machine learning experts, and medical trainees. Through discussion and consensus, the panel categorized symptoms as *present*, *absent*, or *context-dependent* and operationalized their definitions. *Present* indicated symptoms the clinician assessed and the patient endorsed (e.g. *chest pain, fatigue, palpitations, or vomiting*). For *present* symptoms, annotators also coded *modifiers*, words or terms addressing a symptom's frequency or severity (e.g. *controlled, decreased, extreme, frequent, reduced, or stable*). *Absent* indicated symptoms that the clinician assessed but the patient did not endorse (e.g. *denied chest pain, denied shortness of breath*). Each *absent* symptom was also annotated with a corresponding *negation*, the specific word/term indicating the absence of that symptom (e.g. *denies, free, resolved, without*).

Context-dependent indicated any symptom describing "homeostatic" physiological functions such as appetite, sleep, or urination (e.g. *decreased sleep, increased urination*) that

may or may not be pathologic depending on the patient's baseline condition. For example, *weight gain* could indicate either fluid accumulation due to worsening heart failure or an improvement in appetite due to decreased gut edema associated with a higher dose of diuretics. *Context-dependent* also included New York Heart Association (NYHA) heart failure symptom status (e.g. *NYHA class II-III, CHF class II*).

Once symptoms were defined, three members of the study team (*annotators*) reviewed each note in the dataset. Only the HPI and ROS sections of the note were coded. We did not remove other sections of the notes through preprocessing as we sought to evaluate the model's performance on notes with minimal human intervention.

After the annotators each coded 50 notes, a fourth investigator reviewed the annotations and resolved discrepancies. The PyCCI software allowed the reviewer to see words with coding concordance (highlighted in green) or discordance (highlighted in red). If the reviewer was unsure of how to resolve discordant coding, the study PI also provided input. The reviewer gave annotators feedback on areas of common disagreement, which they resolved through discussion and, when necessary, formalized into iterative coding "rules." Annotators repeated this process for each set of 50 notes. To account for the significant variability in how clinicians describe symptoms in free-text notes, the annotators and reviewer classified each annotated symptom based on common terminology (e.g. *chest pain*, *abdominal pain*, *headache)*, used for the ROS in Epic® (Epic Systems, Inc), the most widely used her in the US.

Algorithm

We adapted a deep NLP algorithm, GraphIE(36), to the clinical setting according to the standard three-step model of training, validation, and testing. While many NLP algorithms

analyze the precise sequence of words in a text by abstracting associations between words based on their distance from one another, GraphIE utilizes a graphical structure to abstract relations between words which may exist beyond their sequential positioning. This allows us to analyze both sentence-level patterns as well as more global, non-sequential patterns in the spontaneous language used to describe patient symptoms in clinical notes.

Briefly, the algorithm encodes each clinical note as a set of sentences, and the algorithm labels words as being associated or unassociated, with phrases indicating symptoms. Each note is thus represented by its own graphical structure, with each word being a node, or point, on the graph, and the words' associations with one another represented as edges, or connections, between the word nodes.

Once the graph is built, the algorithm is trained to associate both local sequential and non-local coreferential dependencies between the words and the symptoms to which the words may refer. These graphs are then decoded back to their spontaneously written text-based format, with word labeling derived from the characteristics of the graphical representation. Code and further documentation are available at https://github.com/thomas0809/GraphIE

2.2 Data Analysis

Clinical notes

We quantitatively describe the clinical notes, providing: 1) information about their length and vocabulary variance; 2) the average number per note of each symptom type; 3) the average character and word length of each symptom type, and the number of the related modifiers. We also calculated the symptom distribution for each category.

Evaluation and Metrics

Following a common procedure in the NLP literature, we split the dataset in 80% clinical notes for training (*Train*), 10% for development (*Dev*) and 10% for testing (*Test*)(30,37–42). We calculated the model's performance based on the test set and report accuracy, precision, recall, and the F1 score, which is the harmonic mean of precision and recall and scores range from 0 (worst) to 1 (perfect precision and recall). We calculated these metrics according to standard formulas described in the literature(30). All data analysis was performed using Python version 3.5.

3.0 Results

Patient and Note Characteristics

The total dataset includes 10,870 notes for 990 unique patients, from which we randomly extracted and annotated 154 discharge summaries for 115 patients, who we describe in Table 1. The admission date for all hospitalizations summarized in the selected notes predates CRT-implantation. Mean patient age was 71.5 years (SD 12.2), 93 (81%) were male, and 99 (86.1%) were White. Mean pre-CRT LVEF was 25.1% (SD 7.5) and for those patients with documented NYHA functional classification information (n=49), most (36, 73.5%) had NYHA class III symptoms.

We describe the characteristics of analyzed notes and most common annotated symptoms in Table 2. Notes were written by clinicians on the patient treatment team (e.g., physicians, nurse practitioners). Investigators annotated the HPI and ROS sections of discharge summaries, both of which the EHR automatically inserted into the discharge summary from the admission's initial

History and Physical (H&P) note. Mean note length was 2,286 words, of which 882 words were unique. Mean number of symptoms per note was: 8.7 (SD 8.4) *present*, 8.4 *absent* (SD 11.5), and 1.4 *context-dependent* (SD 1.9). Documented length was similar for *present* and *absent* symptoms (mean 1.61 words, or 10.8 characters), while it was larger for *context-dependent* symptoms (mean 3.0 words, or 17.5 characters). Most commonly documented *present* symptoms were "chest pain," "pain," "shortness of breath," "dyspnea," and "fatigue." Most commonly documented negative symptoms were "chest pain," "palpitations," "chills," "shortness of breath," and "abdominal pain." Most commonly documented *context-dependent* symptoms were "weight gain," "weight loss," "decreased appetite," "poor PO intake," "NYHA Class II symptoms," and "NYHA Class III symptoms." The distribution of annotated symptoms is depicted in Figure 1.

Evaluation and Metrics

The training set contained 124 notes, with 280,178 total words (Supplemental Table 1). Of these, 1,094 (0.4%) were *present* symptoms, 169 (0.1%) were *absent* symptoms and 1,024 (0.4%) were context-dependent symptoms. The development set contained 15 notes with 36,512 words, of which 157 (0.4%) were *present* symptoms, 35 (0.1%) were *absent* symptoms and 138 (0.4%) were *context-dependent* symptoms. Finally, the test set contained 15 notes with 35,467 words, of which 89 (0.3%) were *present* symptoms, 17 (<0.1%) were *absent* symptoms, and 138 (0.4%) were context-dependent symptoms.

We describe the performance of GraphIE on the test set in Table 3. The algorithm's accuracy was 99.4% and it identified 163/244 symptoms (66.8% recall), among the 35,467 words in the test set. Precision was 77.6%, with 47 words incorrectly tagged as symptoms. The

calculated F1 score was 71.8. The algorithm performed better on *present* (F1 score 70.8) and *absent* (F1 74.7) symptoms than it did on *context-dependent* symptoms (F1 48.3). The ratio of False Positive to True Positive was higher for *context-dependent* symptoms (0.7), than it was for *present* (0.3) and *absent* (0.3) ones, due to lower precision (58.6%) and recall (41.2%) for the *context-dependent* category.

We categorized the algorithm's errors and present them in Table 4 . In type 1 errors (*False Positives*), the algorithm tagged words as symptoms that were not annotated in the gold standard. In most of these errors (Supplemental Table 2), the system detected words that indicate symptoms (e.g. *shortness of breath*) but appeared in sections that the annotators ignored (e.g. Patient Instructions). Type 2 errors (*Phrase Boundaries*) are errors in which the system and the annotators disagreed on the boundaries of the symptom (e.g. whether including the preposition *on* inside or outside the annotated phrase). Type 3 errors (*Inconsistent Expressions)* are symptoms with phrasings that were absent from the training set but appear in the test set (e.g. *walk no more than 20-25 feet on flat ground*). These errors included symptoms with typographical or spelling errors. Type 4 errors are *Categorization Errors* (e.g. *weight loss*, which was annotated as context-dependent, but the algorithm tagged as absent).

4.0 Discussion

We developed and tested a deep NLP algorithm to identify symptoms in a cohort of patients with HFrEF undergoing CRT. With respect to existing artificial intelligence literature, the current investigation: 1) reports on the development and testing of an algorithm to capture symptoms in HFrEF patients; and 2) demonstrates the prevalence of symptoms in heart failure

patients across two academic medical centers using a deep NLP approach. The algorithm learned on a training data set, which was manually annotated by three clinicians and reviewed by a fourth one. Annotators triple annotated and reviewed 154 clinical notes (which included 352,157 words), identifying 1340 present symptoms, 1300 absent symptoms, and 221 context-dependent symptoms, representing 80/103 (78%) of the symptoms in Epic's ROS. When compared to manual coding, the gold-standard, our model identified 64% of present, 71% of absent, and 41% of context-dependent symptoms, with a precision of 79%, 79%, and 58%, respectively.

Precision, recall, and F1 scores were on par with established standards in the literature (37,38,41) despite the challenge of identifying symptoms only in specific sections of clinical notes. We deliberately avoided preprocessing notes in the dataset to evaluate the model's performance on notes where there was no human intervention beyond the gold standard annotations. However, GraphIE is designed to utilize long-distance contextual information when making decisions and relies on the notion that identical words should have the same tag. For example, when *headache* appeared in the HPI and was annotated as a present symptom, GraphIE also annotated it as a present symptom in other sections of the note (such as Assessment and Plan), thereby negatively impacting the system's precision. This specific problem explained most false positives. Moving forward, we plan to eliminate irrelevant sections of the notes prior to running the model, which we expect will significantly improve its performance.

Despite this challenge, our model's performance is promising. The model analyzed notes at the word level, which has an extremely small margin for error. When the model missed a single word in a multi-word symptom, the entire phrase became a false negative. These *phrase boundary* errors were particularly common for context-dependent symptoms. Furthermore, we suspect the sensitivity was low due to difficulties the model experienced in identifying poorly

described or infrequently documented symptoms. For example, “shortness of breath,” a positive symptom, tended to appear the same way across notes, whereas “cold sensitivity” could appear as “cold sensitivity,” “coldness,” or “coolness,” among others. Again, these *categorization errors* were particularly evident with context-dependent symptoms, which were relatively rare. However, the high precision means symptoms identified by the deep NLP model matched those identified by the human coders.

Previous research has shown the promise of artificial intelligence methods to leverage big data to assist with radiologic and pathologic diagnoses, sequence complex genetic codes, and identify the etiology of hospital- and nonhospital bacterial isolates(33,43–45). Our group has shown that similar deep NLP techniques can be used to identify symptom burden and palliative care processes in other seriously ill populations, such as patients with cancer and those requiring emergency surgery(30,40,42). The method works approximately 12,000 times faster at identifying symptoms in the EHR than do human curators, which increases the feasibility of using EHR data for discovery. An efficient and accurate method to identify symptoms, researchers will now be able to study the algorithm’s effectiveness in assisting with CRT-related decision-making related to patient-selection and timing of the procedure. If successful, the algorithm would provide individual clinicians with a structured, reproducible, and data-driven approach to determine whether and when to refer patients for CRT.

To date, we are unaware of published reports on the use of deep NLP algorithms to identify symptoms in HFrEF patients undergoing CRT. Cardiologists, palliative care specialists, and patients identify improving our understanding of the prevalence, impact, and management of CHF-specific symptoms as major research and clinical priorities (7,46). Despite a clear need, little data currently exist. The body of literature that attempts to identify symptoms in HFrEF has

relied on PROs like the MLHFQ and KCCQ(1,28,46). These tools are well-validated, but highly symptomatic patients may find them too burdensome to regularly complete in the clinical setting. Missing data compromises the utility of PROs as a strategy to comprehensively identify when clinical deterioration has crossed a threshold such that referral for CRT is appropriate. Our study demonstrates that the use of machine learning and NLP techniques represent one potential way of overcoming this barrier, particularly as EHR use increases at the healthcare systems level(47,48). Deep NLP methods like GraphIE could also be used in conjunction with PROs to provide information about symptom burden for even the sickest patients. Furthermore, these methods may have broad applicability for palliative care clinicians outside of the HFrEF population to assist with symptom identification and management and quality assurance at the programmatic level.

In the future, deep NLP identification of HFrEF-related symptoms may allow for the structured inclusion of vast amounts of free-text data into algorithms that inform decision-making. Invasive procedures such as CRT implantation, ventricular assist device (VAD) placement, and cardiac transplantation are expensive(49–52), involve significant risk to patients(20,21,23–25,53), and do not benefit all patients who undergo them(20,21,53,54). For CRT, particularly CRT-D, patients are at risk of bleeding, local and systemic infection, and inappropriate defibrillation(26,55); many of these risks are compounded given the need for repeated battery exchanges(56). Informed consent discussions around these procedures can be complex(57) and to facilitate shared decision-making, cardiologists and palliative care clinicians should be able to provide interested patients and families with an estimation of the potential benefits of CRT, not only regarding mortality, but also related to symptom burden. In summary, our deep NLP algorithm has potential to identify which patients may benefit most from CRT

implantation and when, based on symptom trajectories, clinicians should begin to consider the procedure.

Our study has several limitations. First, the GraphIE algorithm can only identify symptoms that clinicians document in the EHR. As such, it will miss symptoms that are systematically underreported, such as depression and anxiety(58,59). Additionally, clinicians document symptoms inconsistently, often using words or phrases that may be idiosyncratic (e.g. "walk no more than 20-25 feet on flat ground without dyspnea" instead of dyspnea on exertion) or introducing typographical errors (e.g. "parathesias" instead of "paresthesia"). While algorithms may capture some of these, they may miss others that do not appear in the training set. Furthermore, the patient population was drawn from a single healthcare system with the racial and ethnic composition of New England. Disparities in the assessment and management of symptoms in patients from ethnic and racial minority populations are well documented(60). Another limitation of our study is that we ran the algorithm on a relatively small sample of notes (n=154). However, GraphIE's unit of analysis is the word, thus the sample size (n=352,157 words) in relation to existing NLP literature was actually quite large (30–33). Finally, although our deep NLP algorithm should be generalizable to other settings, it will never capture symptoms that are not documented. It is therefore critical to ensure that clinicians and healthcare systems have robust and equitable methods to ensure that they assess all patients for symptoms.

5.0 Conclusions

A deep NLP algorithm can be trained to efficiently capture symptoms in patients with HFrEF undergoing CRT with satisfactory precision and recall. Future research is necessary to

study clinical applications of this methodology, which may include identifying patients most and least likely to benefit from future invasive procedures.

6.0 Disclosures

All authors report no conflicts of interest.

7.0 Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

8.0 References

1. Bekelman DB, Rumsfeld JS, Havranek EP, et al. Symptom Burden, Depression, and Spiritual Well-Being: A Comparison of Heart Failure and Advanced Cancer Patients. J Gen Intern Med 2009;24:592–8.
2. Solano JP, Gomes B, Higginson IJ. A comparison of symptom prevalence in far advanced cancer, AIDS, heart disease, chronic obstructive pulmonary disease and renal disease. J Pain Symptom Manage 2006;31:58–69.
3. Goldstein NE, Mather H, McKendrick K, et al. Improving Communication in Heart Failure Patient Care. J Am Coll Cardiol. 2019;74:1682–92.
4. Diop MS, Rudolph JL, Zimmerman KM, Richter MA, Skarf LM. Palliative Care Interventions for Patients with Heart Failure: A Systematic Review and Meta-Analysis. J Palliat. Med. 2017;20:84-92.
5. Chuzi S, Pak ES, Desai AS, Schaefer KG, Warraich HJ. Role of Palliative Care in the Outpatient Management of the Chronic Heart Failure Patient. Curr. Heart Fail. Rep. 2019;16:220-8.
6. O'Donnell AE, Schaefer KG, Stevenson LW, et al. Social Worker–Aided Palliative Care Intervention in High-risk Patients With Heart Failure (SWAP-HF). JAMA Cardiol. 2018;3:516-519.
7. Gelfman LP, Bakitas M, Warner Stevenson L, et al. The State of the Science on Integrating Palliative Care in Heart Failure. J Palliat Med. 2017;20:592-603.
8. Leclercq C, Kass DA. Retiming the failing heart: principles and current clinical status of cardiac resynchronization. J Am Coll Cardiol. 2002;39:194–201.
9. Leclercq C, Hare JM. Ventricular Resynchronization. Circulation. 2004;109:296–9.
10. Nam E, Jr FT, Tracy CM, et al. ACCF/AHA/HRS Focused Update ACCF/AHA/HRS Focused Update of the 2008 Guidelines for Device-Based Therapy of Cardiac Rhythm Abnormalities A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society 2012;126:1784–800.
11. Normand C, Linde C, Singh J, Dickstein K. Indications for Cardiac Resynchronization Therapy. JACC Hear Fail 2018;6:308–16.
12. Vaillant C, Martins RP, Donal E, et al. Resolution of left bundle branch block-induced cardiomyopathy by cardiac resynchronization therapy. J Am Coll Cardiol 2013;61:1089–95.
13. Iuliano S, Fisher SG, Karasik PE, Fletcher RD, Singh SN, Department of Veterans Affairs Survival Trial of Antiarrhythmic Therapy in Congestive Heart Failure. QRS duration and mortality in patients with congestive heart failure. Am Heart J 2002;143:1085–91.
14. Yu C-M, Bleeker GB, Fung JW-H, et al. Left ventricular reverse remodeling but not clinical improvement predicts long-term survival after cardiac resynchronization therapy. Circulation 2005;112:1580–6.
15. Abraham WT, Fisher WG, Smith AL, et al. Cardiac resynchronization in chronic heart failure. N Engl J Med 2002;346:1845–53.
16. Cazeau S, Leclercq C, Lavergne T, et al. Effects of multisite biventricular pacing in patients with heart failure and intraventricular conduction delay. N Engl J Med 2001 Mar 22;344:873–80.
17. McAlister FA, Ezekowitz J, Hooton N, et al. Cardiac resynchronization therapy for

patients with left ventricular systolic dysfunction: a systematic review. JAMA 2007 Jun 13;297:2502–14.
18. Auricchio A, Metra M, Gasparini M, et al. Long-Term Survival of Patients With Heart Failure and Ventricular Conduction Delay Treated With Cardiac Resynchronization Therapy. Am J Cardiol 2007;99:232–8.
19. Moss AJ, Hall WJ, Cannom DS, et al. Cardiac-Resynchronization Therapy for the Prevention of Heart-Failure Events. N Engl J Med 2009;361:1329–38.
20. Bristow MR, Saxon LA, Boehmer J, et al. Cardiac-Resynchronization Therapy with or without an Implantable Defibrillator in Advanced Chronic Heart Failure. N Engl J Med 2004;350:2140–50.
21. Cleland JGF, Daubert J-C, Erdmann E, et al. The effect of cardiac resynchronization on morbidity and mortality in heart failure. N Engl J Med 2005;352:1539–49.
22. Young JB, Abraham WT, Smith AL, et al. Combined Cardiac Resynchronization and Implantable Cardioversion Defibrillation in Advanced Chronic Heart Failure: The MIRACLE ICD Trial. J Am Med Assoc 2003;289:2685–94.
23. Swindle JP, Rich MW, McCann P, Burroughs TE, Hauptman PJ. Implantable cardiac device procedures in older patients: Use and in-hospital outcomes. Arch Intern Med 2010;170:631–7.
24. León AR, Abraham WT, Curtis AB, et al. Safety of transvenous cardiac resynchronization system implantation in patients with chronic heart failure: Combined results of over 2,000 patients from a multicenter study program. J Am Coll Cardiol 2005;46:2348–56.
25. Topkara VK, Kondareddy S, Malik F, Wang IW, Mann DL, Ewald GA, et al. Infectious complications in patients with left ventricular assist device: Etiology and outcomes in the continuous-flow era. Ann Thorac Surg 2010;90:1270–7.
26. Atwater BD, Daubert JP. Implantable cardioverter defibrillators: Risks accompany the life-saving benefits. Heart 2012;98:764–72.
27. Alpert CM, Smith MA, Hummel SL, Hummel EK. Symptom burden in heart failure: assessment, impact on outcomes, and management. Heart Fail Rev 2017;22:25–39.
28. Thompson LE, Bekelman DB, Allen LA, Peterson PN. Patient-Reported Outcomes in Heart Failure: Existing Measures and Future Uses. Curr Heart Fail Rep 2015;12:236–46.
29. Kelkar AA, Spertus J, Pang P, et al. Utility of Patient-Reported Outcome Instruments in Heart Failure. JACC Hear Fail 2016;4:165–75.
30. Forsyth AW, Barzilay R, Hughes KS, et al. Machine Learning Methods to Extract Documentation of Breast Cancer Symptoms From Electronic Health Records. J Pain Symptom Manage 2018;55:1492–9.
31. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Informatics Assoc 2019;26(4):364–79.
32. Sheikhalishahi S, Miotto R, Dudley JT, et al. Natural Language Processing of clinical notes: A systematic review for Chronic Diseases (Preprint). JMIR Med Informatics 2018;7(2):e12239.
33. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017;73:14–29.
34. Murphy SN, Gainer V, Chueh HC. A visual interface designed for novice users to find research patient cohorts in a large biomedical database. AMIA . Annu Symp proceedings

AMIA Symp 2003;2003:489–93.
35. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a Research Patient Data Repository. AMIA . Annu Symp proceedings AMIA Symp 2006;2006:1044.
36. Qian C, Wen L, Kumar A. TraceWalk: Semantic-based Process Graph Embedding for Consistency Checking. 2019. Available from: http://arxiv.org/abs/1905.06883. Accessed July 19, 2019.
37. Kadra G, Stewart R, Shetty H, et al. Extracting antipsychotic polypharmacy data from electronic health records: Developing and evaluating a novel process. BMC Psychiatry 2015;15:166.
38. Lin C, Karlson EW, Dligach D, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. J Am Med Informatics Assoc 2015;22:e151–61.
39. Chan A, Chien I, Moseley E, et al. Deep learning algorithms to identify documentation of serious illness conversations during intensive care unit admissions. Palliat Med 2019;33:187–96.
40. Lindvall C, Lilley EJ, Zupanc SN, et al. Natural Language Processing to Assess End-of-Life Quality Indicators in Cancer Patients Receiving Palliative Surgery. J Palliat Med;22(2):183–7.
41. Carrell DS, Halgrim S, Tran DT, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. Am J Epidemiol 2014;179:749–58.
42. Udelsman B, Chien I, Ouchi K, et al. Needle in a Haystack: Natural Language Processing to Identify Serious Illness. J Palliat Med. 2018;22:179–82.
43. Senders JT, Staples PC, Karhade A V., et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. World Neurosurg 2018;109:476-486
44. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016;375:1216–9.
45. Tolo I, Thomas JC, Fischer RSB, et al. Do Staphylococcus epidermidis Genetic Clusters Predict Isolation Sources? J Clin Microbiol 2016;54:1711–9.
46. Stevenson LW, Hellkamp AS, Leier C V, et al. Changing preferences for survival after hospitalization with advanced heart failure. J Am Coll Cardiol 2008;52:1702–8.
47. Adler-Milstein J, DesRoches CM, Kralovec P, et al. Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. Health Aff 2015;34:2174–80.
48. Adler-Milstein J, Holmgren AJ, Kralovec P, et al. Electronic health record adoption in US hospitals: the emergence of a digital “advanced use” divide. J Am Med Informatics Assoc 2017;24:1142–8.
49. Shah N, Agarwal V, Patel N, et al. National Trends in Utilization, Mortality, Complications, and Cost of Care after Left Ventricular Assist Device Implantation from 2005 to 2011. Ann Thorac Surg 2016;101:1477–84.
50. Woo CY, Strandberg EJ, Schmiegelow MD, et al. Cost-effectiveness of adding cardiac resynchronization therapy to an implantable cardioverter-defibrillator among patients with mild heart failure. Ann Intern Med 2015;163:417–26.
51. Mealing S, Woods B, Hawkins N, et al. Cost-effectiveness of implantable cardiac devices in patients with systolic heart failure. Heart 2016;102(21):1742–9.
52. Baras Shreibati J, Goldhaber-Fiebert JD, Banerjee D, Owens DK, Hlatky MA. Cost-Effectiveness of Left Ventricular Assist Devices in Ambulatory Patients With Advanced

Heart Failure. JACC Hear Fail 2017;5(2):110–9.
53. McAlister FA, Ezekowitz J, Hooton N, et al. Cardiac resynchronization therapy for patients with left ventricular systolic dysfunction: A systematic review. JAMA 2007;297:2502–14.
54. Nassif ME, Tang Y, Cleland JG, et al. Precision Medicine for Cardiac Resynchronization: Predicting Quality of Life Benefits for Individual Patients-An Analysis from 5 Clinical Trials. Circ Hear Fail 2017;10: pii: e004111.
55. Kipp R, Hsu JC, Freeman J, et al. Long-term morbidity and mortality after implantable cardioverter-defibrillator implantation with procedural complication: A report from the National Cardiovascular Data Registry. Hear Rhythm. 2018;15:847–54.
56. Lewis KB, Stacey D, Carroll SL, et al. Estimating the Risks and Benefits of Implantable Cardioverter Defibrillator Generator Replacement: A Systematic Review. Pacing and Clinical Electrophysiology. 2016;39:709–22.
57. Allen LA, Stevenson LW, Grady KL, et al. Decision making in advanced heart failure: A scientific statement from the american heart association. Circulation 2012;125(15):1928–52.
58. Borowsky SJ, Rubenstein L V., Meredith LS, et al. Who is at risk of nondetection of mental health problems in primary care? J Gen Intern Med 2000;15:381–8.
59. Luoma JB, Martin CE, Pearson JL. Contact with mental health and primary care providers before suicide: A review of the evidence. Am J Psychiatry 2002;159:909–16.
60. Johnson KS. Racial and Ethnic Disparities in Palliative Care. J Palliat Med 2013;16:1329–34.

Table 1. Baseline characteristics of study participants at the time of CRT-implantation

Characteristics, N = 115	
Age, mean years (SD)	71.2 (12.2)
Male gender, n (%)	93 (80.9)
Race	
White	99 (86.1)
Non-White	16 (13.9)
NYHA class[a], n=49 (%)	
I	1 (2.0)
II	12 (24.5)
III	36 (73.5)
IV	0
LVEF, mean % (SD)	25.1 (7.5)
Heart rate, mean BPM (SD)	84.1 (40.0)
QRS duration, mean ms (SD)	158.1 (35.4)
Left bundle branch block, n (%)	39 (33.9)
AV block, n (%)	20 (17.4)
Afib, n (%)	65 (56.5)
Renal disease, n (%)	15 (13.0)
Baseline creatinine, mean mg/dL (SD)	2.3 (2.8)
Diabetes, n (%)	42 (36.5)
CRT response[b], n (%)	
Died within year	34 (29.6)
Non-Responder	17 (14.8)
Responder	64 (55.6)

NYHA= New York Heart Association LVEF= left ventricular ejection fraction

[a] 66 patients missing NYHA functional classification information.

[b] CRT response: Non-responder = <0% improvement in LVEF 6-18 months following CRT, Response = ≥0% improvement in LVEF 6-18 months following CRT

Table 2. Characteristics of notes and investigator-annotated symptoms.

	# reports	Avg. words	Avg. unique words	Avg. Present Symptoms	Avg. Modifiers	Avg. Absent Symptoms	Avg. Negation	Avg. Context-dependent Symptoms
Dataset	154	2286	882	8.7	NA	8.4	NA	1.4
Avg. # words per symptom				1.7	1.3	1.5	1.0	3.0
Avg. # characters of symptoms				11.2	9.9	10.3	4.2	17.5
Top ten frequent symptoms/words				Chest pain Pain Shortness of breath Dyspnea Fatigue Nausea Abdominal Pain Orthopnea Cough Syncope	Worsening increased intermittent progressive severe persistent mild improved chronic more	Chest pain Palpitations Chills Shortness of Breath Abdominal pain Vomiting Nausea Syncope Diarrhea Cough	No denies negative not resolved without free less relief never	Weight gain Weight loss Decreased appetite Poor PO intake NYHA Class II symptoms NYHA class III symptoms congestive heart failure symptoms class IV heart failure Ambulated independently without difficulty Concentrated urine

Table 3. Results of GraphIE on the test set.

	Accuracy	**Annotated Number of Symptoms**	**True Positive**	**False Positive**	**Precision**	**Recall**	**F1**
Present	N/A	89	57	15	79.2%	64.0%	70.8
Absent	N/A	138	99	27	78.6%	71.2%	74.7
Context-dependent	N/A	17	7	5	58.3%	41.2%	48.3
Overall	**99.4%**	**244**	**163**	**47**	**77.6%**	**66.8%**	**71.8**

Table 4. Typology of errors with examples (**Bold** text indicates symptom labeling)

	ANNOTATORS	ALGORITHM
TYPE 1	FALSE POSITIVES	
	You were admitted to the Brigham and Women's Hospital Cardiology Service for shortness of breath	You were admitted to the Brigham and Women 's Hospital Cardiology Service for **shortness of breath**
	Cardiac: SOB, CP, palpitations, dependent edema	**Cardiac: SOB, CP, palpitations, dependent edema**
	He denied any angina, palpitations, syncope, change in diet or change in medications	He **denied** any **angina, palpitations, syncope, change in diet** or **change in medications**
	Patient had repeat evacuation for increasing lethargy and gait instability.	Patient had repeat evacuation for **increasing lethargy** and **gait instability.**
	Patient demonstrated steady gait with decreased step length on RLE and decreased stance time of LLE with straight gait path	Patient demonstrated steady gait with decreased step length on RLE and decreased stance time of **LLE with straight gait path**
TYPE 2	PHRASE BOUNDARIES	
	Patient is **short of breath** on **mild to moderate exertion.**	Patient is **short of breath on mild to moderate exertion.**
	Right sided back **pain.**	**Right sided back pain.**
	Cardiac: SOB, CP, palpitations, dependent edema	**Cardiac: SOB, CP, palpitations, dependent edema**
	He presents with episodes of **exertional chest pain,** associated bilateral arm **numbness**.	He presents with episodes of **exertional chest pain**, associated bilateral **arm numbness.**
TYPE 3	INCONSISTENT EXPRESSIONS	
	He estimates that he could **walk no more than 20-25 feet on flat ground** without **dyspnea.**	He estimates that he could walk no more than 20-25 feet on flat ground without dyspnea.
	He notes that **his feet might have been cold** throughout the summer	He notes that his feet might have been cold throughout the summer
	Patient denies **parathesias** *(sic)*	Patient denies parathesias *(sic)*
	Last episode of **CP** was yesterday evening at 7pm.	Last episode of CP was yesterday evening at 7pm.
TYPE 4	CATEGORIZATION ERRORS	
	Abdominal bloating, loss of appetite (Algorithm: present; Annotators: context-dependent)	**Abdominal bloating, loss of appetite** (Algorithm: present; Annotators: context-dependent)
	Constitutional: **Negative** for **fever**, **chills**, and **weight loss** (Algorithm: absent; Annotators: context-dependent)	Constitutional: **Negative** for **fever**, **chills**, and **weight loss** (Algorithm: absent; Annotators: context-dependent)