SUPPLEMENTAL MATERIAL
SUPPLEMENTAL METHODS

Angina NLP System Overview

To meet the goals of the project, a natural language processing (NLP) system was developed to identify instances of the Canadian Cardiovascular Society (CCS) angina score explicitly stated within clinical notes. All clinical documents available for a cohort of 6,556,919 VA patients were used to train and validate the system. A training set was created for a random set of patients within cohort. This set contained 1,885,840 documents and 7,456 patients. These documents were used to create the knowledge base for the NLP system.

NLP Pipeline Methods

The NLP system is based on the Unstructured Information Management Architecture Asynchronous Scaleout (UIMA AS; Ferrucci & Lally, 2004). The system also utilizes the VINCI Leo framework, and a set of libraries that enables more efficient utilization of UIMA AS (Cornia, et al, 2014).

The NLP pipeline was developed over several iterations that included cycles of system development and error analysis. The system design included several sets of manually created regular expressions in order to identify concepts related to CCS scores. The co-occurrence patterns of words and phrases that were found in text through the use of these regular expressions were used to disambiguate relevant concepts. The NLP system has the following modules:
1. Regular Expressions:

[CCS_Concept] – These expressions are used to identify mentions of CCS score such as “CCS”, “CCSC”, “Canadian”, “Functional class” and “angina”.

[Concept_word] – These expressions enable detection of words that are frequently used in a CCS phrase, such as “grade”, “class”, “score”, “stage”.

[CCS_Score] – These expressions capture numeric values typically associated with a score, such as 1,2,3,4,I,II,III,IV,1-2, 2-3, etc...

[Exclude] – These expressions identify words and numbers that indicate that the surrounding phrase does not describe CCS score, such as dates, other irrelevant numeric values, and invalid units of measure.

2. Patterns:

After all of these various terms and characters are captured, two pattern types are used in order to separate relevant phrases from the irrelevant. These two patterns are CCS_Pattern and Exclude_Pattern.

[CCS_Pattern] – These patterns describe valid sequences of CCS_Concept, Concept_word, and CCS Score that can be encountered in the text. Examples of such sequences are outlined in Supplemental Table 4.

A few very specific patterns were created in order to capture special cases like questionnaires and form, such as “[x] Angina 3?”

[Exclude_Pattern] – These patterns describe sequences of previously captured phrases that indicate that are actually irrelevant and do not actually represent CCS score. For example, a phase “Angina 2x/week” contains CCS_concept phrase "angina" and
CCS_score phrase "2", but this phrase is detected as irrelevant because it also contains an Exclude phrase "x/week", which indicates that the phrase is irrelevant. Other examples of phrases captured by exclusion patterns are “Angina 2/2”, “Angina 2/13/12”, and “1000ccs given to pt”.

SUPPLEMENTAL RESULTS

All clinical notes for all 6,556,919 patients in the cohort were processed. The final output contained 64,732 mentions of CCS scores for 22,994 patients. NLP systems are typically evaluated using precision, recall, and F-1 score performance metrics.

For precision validation, we selected 500 documents that contained CCSscore mentions identified by the NLP system. The validation set had 609 mentions that were manually reviewed. The validation step was completed using the Chex validation tool (DuVall, et al 2014.). This validation found that the NLP system had 93.1% precision (positive predictive value).

Due to the extreme rarity of the CCS concept mentions in clinical notes, an enriched set of notes was needed in order to keep human annotators from reading over 1000 documents before finding a single instance of a CCS score. Sampling was done using a set of filtering criteria. First, the documents were filtered by only using documents from the date of an angina diagnosis, determined by ICD-9 codes. Using this single filter, we found that CCS scores were found in less than 0.5% of all documents. The corpus was further filtered by documents containing “card” in the document title (cardiology, cardiologist, etc.). Lastly, the presence of the word “angina” or “CAD” (coronary artery disease) was required to be in the documents. We were able to estimate that CCS could be found in roughly 6% of the documents after all of these filters were applied.
Using the filtering criteria, 500 documents identified for manual annotation. Three annotators reviewed the documents using the eHost annotation tool. They found 33 instances of a CCS score, 25 of which were also found by the system. This sensitivity analysis found that the NLP system had a 75.7% recall (sensitivity).

In addition to explicit CCS scores, the annotators also marked any instance of angina being described in terms of exertion, which could be used to infer CCS score (i.e. “angina after 2 flights of stairs”, “angina after walking two blocks.”). In the same 500 documents, 240 instances of exertional angina descriptions were found. These statements were not used for system development but can serve as a basis for future work.

Limitations and Future Work

The rarity of the CSS score mentions in clinical notes leads to several limitations:

First, even after being enriched with several filters, the recall corpus of 500 manually reviewed documents found only 33 instances of a CCS score. Such a low number of CCS score mentions increases likelihood that the true recall differs from the measured value.

Second, the set of filters used to select documents for review introduce a risk of selection bias.

Error analysis revealed that a relatively low recall was due to several causes. First, new abbreviations appeared in the selected documents that were not found in the training set. Angina Pectoris was abbreviated twice as AP (i.e. “AP class II”) and the Functional class was abbreviated as FC (i.e. “pt complains of angina FC III”). Another cause for a
lower than expected recall score was the instance of novel spellings that had not been encountered during the knowledge base acquisition. In one instance, “CCC class I” was found rather than the expected CCS.

Future work may include the following steps: 1) With additional effort, another development iteration can be performed to improve the recall and precision. 2) CCS score can be potentially inferred using exertional angina descriptions. The manual annotations performed for sensitivity analysis can be used as a training set for a potential CCS score inference algorithm.
Table S1. Phelp’s Criteria for classification of chronic stable angina. The protocol uses three definitions (Criteria A, B or C) to identify angina veterans from a combination of ICD-9-CM diagnostic and procedural codes, CPT codes, and medication use.

| Criteria | Criteria requirements |
|----------|-----------------------|
| A        | ≥ 2 diagnosis codes for CAD on separate dates within a 12-month period  
          | ≥ 2 diagnosis codes for chest pain on separate dates  
          | ≥ 2 nitrate prescriptions with a gap of at least 30 days between fill dates |
| B        | ≥ 1 diagnosis code for stable angina  
          | ≥ 2 prescription fills for nitrates with a gap of at least 30 days between fills within a 12-month period |
| C        | ≥ 2 diagnosis codes for stable angina on separate dates (one if it is a primary diagnosis during an inpatient stay)  
          | Did not have at least two pharmacy fills for nitrates with a gap of at least 30 days  
          | Either ≥ 2 beta-blocker prescriptions or ≥ 2 calcium channel blocker prescriptions with a gap of at least 30 days between fills within a 12-month period |

| Condition | Code type | Code/Description |
|-----------|-----------|------------------|
| Chest pain| ICD-9-CM diagnosis | 786.5 (Chest pain, unspecified), 786.51 (Precordial pain), 786.52 (Painful respiration), 786.59 (Other chest pain) |
| Stable angina | ICD-9-CM diagnosis | 413.x (Angina pectoris) |
|---------------|--------------------|------------------------|
| **CAD**       | ICD-9-CM diagnosis | 410.xx (Acute myocardial infarction), 411.xx (Other acute and sub-acute forms of ischemic heart disease), 412 (Old myocardial infarction), 414.xx (Other forms of chronic ischemic heart disease), 429.2 (Cardiovascular disease, unspecified), 429.5 (Rupture of chordae tendinae), 429.6 (Rupture of papillary muscle), 429.7x (Certain sequelae of myocardial infarction, not elsewhere specified), 996.03 (Mechanical complications due to coronary bypass graft), V45.81 (Aortocoronary bypass status), V45.52 (Percutaneous transluminal coronary angioplasty status) |
| ICD-9-CM procedure | 0.66 (Percutaneous transluminal coronary angioplasty or atherectomy), 36.0x (Removal of coronary artery obstruction and insertion of stent), 36.1x (Bypass anastomosis for heart revascularization), 36.2 (Heart revascularization by arterial implant), 36.3x (Other heart revascularization) |
| CPT          | Description                                                                 |
|--------------|-----------------------------------------------------------------------------|
| 33140-33141  | (Transmyocardial laser revascularization, by thoracotomy),                 |
|              | 33510-33523, 33533-33536 (Coronary artery bypass),                         |
|              | 33572 (Coronary endarterectomy in conjunction with CABG),                   |
|              | 92975 (Thrombolysis, coronary; by intracoronary infusion),                  |
|              | 92980-92981 (Transcatheter placement of an intracoronary stent, percutaneous, with or without other therapeutic intervention, any method), 92982, 92984 (Percutaneous transluminal coronary balloon angioplasty), 92995-92996 (Percutaneous transluminal coronary atherectomy, by mechanical or other method, with or without balloon angioplasty), 93540 (Injection procedure during cardiac catheterization for selective opacification of aortocoronary venous bypass grafts) |
| HCPCS        | G0290-G0291 (Transcatheter placement of drug eluting intracoronary stent),  |
|              | S0340-S0342 (Lifestyle modification program for management of CAD), S2205-S2209 (Minimally invasive direct coronary artery bypass surgery involving mini-thoracotomy or mini-sternotomy surgery using arterial and/or venous graft), G8033-G8041, G8498 (Patient coded as a coronary artery disease patient), G8159-G8172, G9497 (Patient coded as a coronary artery bypass graft patient) |
CABG = Coronary artery bypass grafting; CAD = Coronary artery disease; CPT = Current Procedural Terminology; HCPCS = Healthcare Common Procedure Coding System; ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification

*Adapted with permission from Phelps CE, Buysman EK, Gomez Rey G. Costs and clinical outcomes associated with use of ranolazine for treatment of angina. Clin Ther. 2012;34:1395–1407.e4.
Table S2. Examples of used patterns and phrases that match the patterns

| Pattern                                      | Example                        |
|----------------------------------------------|--------------------------------|
| [CCS_word] [CCS_score] [CCS_Concept]         | “class IV angina”              |
|                                              | “grade 2 CCS”                  |
| [CC_Concept] [CCS_Word] [CCS_score]          | “Angina class 2”               |
|                                              | “Canadian score 3”             |
| [CCS_concept] [CCS_score]                    | "CCS 1"                        |
|                                              | "functional class 3"           |
|                                              | "angina IV"                    |
Table S3. Codes used to define variables in the analysis

| Variable                        | Code(s)                                                                 |
|---------------------------------|-------------------------------------------------------------------------|
| Angina                          | ≥1 claim with an ICD-9-CM diagnosis (any position) of 413.x.            |
| Chest pain                      | ≥1 claim with an ICD-9-CM diagnosis (any position) of 786.50, 786.51, 786.52, and 786.59. |
| Coronary artery disease         | Any one of the following:                                               |
|                                 | a) ≥1 claim with an ICD-9-CM diagnosis (any position) of 410.xx, 411.x, 412, 414.xx, 429.2, 429.5, 429.6, 429.7x, 996.03, V45.81, V45.82 |
|                                 | b) ≥1 claim with an ICD-9-CM procedure (any position) of 00.66, 36.0x, 36.1x, 36.2, 36.3x |
|                                 | c) ≥1 claim with a CPT code (any position) of 33140, 33141, 33510-33523, 33533-33536, 33572, 92975, 9280, 92981, 92982, 92984,92995, 92996, 93540 |
|                                 | d) ≥1 claim with a HCPCS code (any position) of G0290, G0291, S0340-S0342, S2205-S2209, G8033-G8041, G8498, G8159-G8172, G9497 |

Baseline characteristics
| Condition                  | Definition                                                                 |
|----------------------------|---------------------------------------------------------------------------|
| Dyslipidemia               | ≥1 claim with an ICD-9-CM diagnosis (any position) of 272.4                |
| Diabetes mellitus          | ≥1 claim with an ICD-9-CM diagnosis (any position) of 249.xx, 250.xx, 357.2, 362.0x, 366.41 |
| Myocardial infarction      | Any one of the following:                                                 |
|                            | a) Acute myocardial infarction: ≥1 claim with an ICD-9-CM diagnosis (any position) of 410.xx, 412 |
|                            | b) History of myocardial infarction: ≥1 claim with an ICD-9-CM diagnosis (any position) of 411.0, 412 |
| Hypertension               | ≥1 claim with an ICD-9-CM diagnosis (any position) of 401.xx, 402.xx, 403.xx, 404.xx, 405.xx |
| Heart failure              | ≥1 claim with an ICD-9-CM diagnosis (any position) of 398.91, 402.01, 402.11, 402.91 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.xx |
| Smoking                    | ≥1 claim with an ICD-9-CM diagnosis (any position) of 305.1, 649.0x, 989.84, V15.82 |
| Atrial fibrillation        | ≥1 claim with an ICD-9-CM diagnosis (any position) of 427.31               |
| Healthcare utilization metrics          |
|----------------------------------------|
| **6-month pre-index costs**            |
| Sum of the inpatient and outpatient total costs. Use the variable `acttotcost` from DSS Outpatient for outpatient costs and the variable `totcostin` from the DSS Discharge dataset for inpatient costs. |

| **6-month pre-index ED visits**        |
| Identified by primary or secondary clinic stop codes 101 (prior to 2007) and 130 and 131 (since 2007). Multiple ED codes on the same service date counted as one visit. |

| **6-month pre-index hospitalizations** |
| Numerical variable of the number of hospitalizations by the patient in the 6-month pre-index period. |

| **Total costs**                        |
| Sum of the inpatient and outpatient total costs. Use the variable `acttotcost` from DSS Outpatient for outpatient costs and the variable `totcostin` from the DSS Discharge dataset for inpatient costs. |

| **Outpatient costs**                   |
| Sum of the outpatient total costs using the variable `acttotcost` from DSS Outpatient. |

| **Inpatient costs**                    |
| Sum of the inpatient total costs using the variable `totcostin` from DSS Discharge dataset. |

| **Pharmacy costs**                     |
| Sum of the total pharmacy costs using DSS. |
| Procedure                                                   | Any one of the following:                                                                                                                                 |
|-------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| Percutaneous coronary intervention                          | a) ≥1 claim with a CPT code (any position) of 92980, 92981, 92982, 92984, 92995, 92996, 92920, 92921, 92924, 92925, 92928, 92929, 92933, 92934, 92937, 92938, 92941, 92943, 92944, 92973, 92973, 92975, 92978, 92979.  
   b) ≥1 claim with a HCPCS code (any position) of G0290-G0291, C9600-C9608  
   c) ≥1 claim with an ICD-9-CM procedure code (any position) of 00.66, 36.01, 36.02, 36.05, 36.06, 36.07                                                                 |
| Coronary artery bypass graft                                | a) ≥1 claim with a CPT code (any position) of 33510-33516, 33517-33523, 33533-33536  
   b) ≥1 claim with an HCPCS code (any position) of S2205-S2209  
   c) ≥1 claim with an ICD-9-CM procedure code (any position) of 36.1x.                                                                 |
| Emergent percutaneous coronary intervention                 | A percutaneous coronary intervention with a corresponding hospitalization and ICD-9-CM of 410.x. This method has been previously been validated. (30) }
| All-cause hospitalizations | Identified as an inpatient visit (Medical or Surgical) in the CDW. |
|---------------------------|---------------------------------------------------------------|
| Acute coronary syndrome hospitalizations | ≥1 claim with an ICD-9-CM diagnosis (any position) of 410.00-410.92 associated with an inpatient visit |
| Heart failure hospitalizations | ≥1 claim with an ICD-9-CM diagnosis (any position) of 398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.0-428.9 associated with an inpatient visit |
| Atrial fibrillation hospitalizations | ≥1 claim with an ICD-9-CM diagnosis (any position) of 427.31 associated with an inpatient visit |
| Emergency department visits | Post-index period ED visits were identified by primary or secondary clinic stop codes 101 (prior to 2007) and 130 and 131 (since 2007) in the CDW and counted. Multiple ED codes on the same service date were counted as one visit. |
| Outpatient visits | Outpatient visits were identified by primary or secondary clinic stop codes in the CDW and included all outpatient visits including physician visits, laboratory, radiology, physical therapy, etc. |
| Category                      | Description                                                                                                                                 |
|-------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| General medicine outpatient   | Outpatient visits were identified by primary or secondary clinic stop codes 301, 323, 348, or 27 in the CDW and included all outpatient visits including physician visits, laboratory, radiology, physical therapy, etc. |
| Cardiology outpatient visits  | Cardiology Outpatient visits were identified by primary or secondary clinic stop codes 303 in the CDW and included all outpatient visits including physician visits, laboratory, radiology, physical therapy, etc. |

*CDW = Corporate Data Warehouse; CPT = Current Procedural Terminology; DSS = Decision Support System; HCPCS = Healthcare Common Procedure Coding System; ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification.*
Table S4. Characteristics of Angina Patients With and Without a CCS Class Documentation

| Characteristic       | Without (n=276,600) | With (n=14,216) | p Value |
|----------------------|---------------------|-----------------|---------|
| **Age, y**           | 67 ± 11             | 67 ± 9.8        | <.001   |
| **Male Sex**         | 268877 (97%)        | 13996 (98%)     | <.001   |
| **Race**             |                     |                 |         |
| White                | 214867 (78%)        | 11464 (81%)     | <.001   |
| Black                | 35765 (13%)         | 1662 (12%)      | <.001   |
| Asian                | 3721 (1%)           | 240 (2%)        | <.001   |
| American Indian      | 2542 (1%)           | 135 (1%)        | 0.705   |
| Unknown/Missing      | 19705 (7%)          | 715 (5%)        | <.001   |
| **Census Region**    |                     |                 |         |
| Northeast            | 31469               | 1387            | <.001   |
| Midwest              | 65717               | 2240            | <.001   |
| South                | 125785              | 6948            | <.001   |
| West                 | 46683               | 2365            | 0.454   |
| Missing/Other        | 6946                | 1276            | <.001   |
| **Pre-index comorbidities and clinical characteristics** | | | |
| CCI Score            | 1.3 ± 2.2           | 1.1 ± 2.0       | <.001   |
| Dyslipidemia         | 153955 (56%)        | 8670 (61%)      | <.001   |
| Condition                        | Control  | Case       | p-Value |
|--------------------------------|----------|------------|---------|
| Diabetes Mellitus               | 114191 (41%) | 6650 (47%) | <.001   |
| Acute or history of MI          | 21138 (8%) | 0 (0%)     |         |
| Hypertension                    | 204123 (74%) | 11380 (80%) | <.001 |
| Heart Failure                   | 51430 (19%) | 2677 (19%) | 0.476 |
| Smoking                         | 60761 (22%) | 2949 (21%) | <.001 |
| Alcohol Abuse                   | 19407 (7%) | 838 (6%)   | <.001   |
| Body Mass Index                 | 30 ± 6.3  | 31 ± 5.9   | <.001   |
| **BMI Categories**              |          |            |         |
| Underweight                     | 2901 (1%) | 98 (1%)    | <.001   |
| Normal                          | 49878 (18%) | 2158 (15%) | <.001   |
| Overweight                      | 95688 (35%) | 5083 (36%) | 0.006   |
| Obese                           | 127875 (46%) | 6877 (48%) | <.001   |
| **Medications**                 |          |            |         |
| anti-Anginals                   | 1.3 ± 0.80 | 1.3 ± 0.86 | <.001   |
| Anti-Anginals                   |          |            |         |
| Beta-blocker                    | 196028 (71%) | 10233 (72%) | 0.005   |
| Calcium channel blocker         | 62905 (23%) | 3467 (24%) | <.001   |
| Long Acting Nitrate             | 96800 (35%) | 5354 (38%) | <.001   |
| Aspirin                         | 95882 (35%) | 4929 (35%) |         |
| Statin                          | 198773 (72%) | 10196 (72%) | 0.715   |
| ACE Inhibitor or ARB            | 164056 (59%) | 8665 (61%) | <.001   |
| P2Y12 Antagonist                | 67037 (24%) | 3226 (23%) | <.001   |
|                  | Yes          | No           | P value |
|------------------|--------------|--------------|---------|
| Patients having  | 110392 (40%) | 122813 (44%) | <.001   |
| Number of        |              |              |         |
| Hospitalizations | 7438 (52%)   | 4462 (31%)   | <.001   |
|                 |              |              |         |
| Unknown/Missing  | 43395 (16%)  | 4462 (31%)   | 0.051   |
|                 |              |              |         |
| Seeing a         | 173269 (63%) | 122813 (44%) | <.001   |
| cardiologist     | 7975 (56%)   | 4462 (31%)   |         |

**Preindex Resource**

**Utilization and Costs**

**Hospitalizations**

|                  | Yes          | No           | P value |
|------------------|--------------|--------------|---------|
| Patients having  | 85243 (31%)  | 3552 (25%)   | <.001   |
| Number of        | 0.54 ± 1.1   | 0.37 ± 0.78  | <.001   |
| Hospitalizations |              |              |         |

**ER events**

|                  | Yes          | No           | P value |
|------------------|--------------|--------------|---------|
| Patients Having  | 107567 (39%) | 8202 (58%)   | <.001   |
| Number of Events | 0.96 ± 1.9   | 3.7 ± 7.2    | <.001   |
| Total Costs      | 16221 ± 40274| 35148 ± 95986| <.001   |

Data are number of patients (%) or mean (SD). CCI: Charlson/Quan Comorbidity Index, PCI=percutaneous coronary intervention, CABG=coronary artery bypass graft, CCB=calcium channel blocker, ACE=Angiotensin converting enzyme, ARB=Angiotensin II Receptor Blocker
Supplemental References:

Cornia R, Patterson O V, Ginter T, Duvall SL. Rapid NLP Development with Leo. In: AMIA Annu Symp Proc.; 2014.

Duvall SL, Cornia RC, Forbush TB, Halls CH, Patterson O V. Check it with Chex : A Validation Tool for Iterative NLP Development. In: AMIA Annu Symp Proc.; 2014.

Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering, 10(3-4).

South, B. R., Leng, J., Anderson, K., Shen, S., Thibault, J., & DuVall, S. (2010). The Extensible Human Oracle Suite of Tools (eHOST) for Pre-Annotation of Clinical Narratives. In BioCreative: Critical Assessment of Information Extraction in Biology Annual Conference.