HHS Public Access

Author manuscript

Kidney Int. Author manuscript; available in PMC 2021 February 01.

Published in final edited form as:

Kidney Int. 2020 February ; 97(2): 383–392. doi:10.1016/j.kint.2019.10.023.

Natural language processing of electronic health records is superior to billing codes to identify symptom burden in hemodialysis patients.

Lili Chan, MD, MS[1,2], **Kelly Beers, DO, MS**[1], **Amy Yau, MD**[1], **Kinsuk Chauhan, MD**[1], **Aine Duffy, MS**[2], **Kumardeep Chaudhary, PhD**[2], **Neha Debnath, MD**[1], **Aparna Saha, MD**[2], **Pattharawin Pattharanitima, MD**[1], **Judy Cho, MD**[2], **Peter Kotanko, MD**[1,3], **Alex Federman, MD**[4], **Steven Coca, DO, MS**[1], **Tielman Van Vleck, PhD**[2], **Girish N. Nadkarni, MD, MPH**[1,2]

[1]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY

[2]The Charles Bronfman Institute for Personalized Medicine, Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY

[3]Renal Research Institute, New York, NY

[4]Division of General Internal Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY

Abstract

Symptoms are common in patients on maintenance hemodialysis but identification is challenging. New informatics approaches including natural language processing (NLP) can be utilized to identify symptoms from narrative clinical documentation. Here we utilized NLP to identify seven patient symptoms from notes of maintenance hemodialysis patients of the Bio*Me* Biobank and validated our findings using a separate cohort and the MIMIC-III database. NLP performance was compared for symptom detection with International Classification of Diseases (ICD)-9/10 codes and the performance of both methods were validated against manual chart review. From 1034 and 519 hemodialysis patients within Bio*Me* and MIMIC-III databases, respectively, the most frequently identified symptoms by NLP were fatigue, pain, and nausea/vomiting. In Bio*Me*, sensitivity for NLP (0.85 – 0.99) was higher than for ICD codes (0.09 – 0.59) for all symptoms with similar results in the Bio*Me* validation cohort and MIMIC-III. ICD codes were significantly more specific for nausea/vomiting in Bio*Me* and more specific for fatigue, depression, and pain in

Correspondence: Lili Chan, MD, MS or Girish N. Nadkarni, MD, MPH, Icahn School of Medicine at Mount Sinai, One Gustave L Levy Place, Box 1243, New York, NY 10029, Telephone number: 212-241-8640 or (212) 241-1385, Fax number: (212) 849-2643, lili.chan@mountsinai.org or girish.nadkarni@mountsinai.org.

Author Contributions: LC, SC, and GNN designed the study. TVV parsed the data. LC, KC, KC and ND carried out the analysis. LC, AY, and KB performed the manual chart review. ND and AS made the figures and tables. All authors drafted and revised the manuscript and approved the final version of the manuscript.

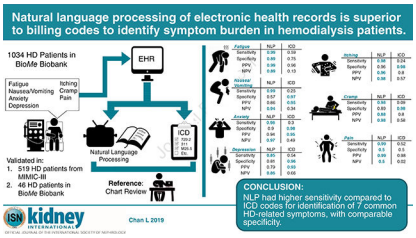
*TVV and GNN contributed equally.

Supplementary Material:

Supplementary information is available at Kidney International's website

the MIMIC-III database. A majority of patients in both cohorts had four or more symptoms. Patients with more symptoms identified by NLP, ICD, and chart review had more clinical encounters. NLP had higher specificity in inpatient notes but higher sensitivity in outpatient notes and performed similarly across pain severity subgroups. Thus, NLP had higher sensitivity compared to ICD codes for identification of seven common hemodialysis-related symptoms, with comparable specificity between the two methods. Hence, NLP may be useful for the high-throughput identification of patient-centered outcomes when using electronic health records.

Graphical Abstract



Keywords

hemodialysis; geriatric nephrology

Introduction:

There are over 450,000 patients on maintenance hemodialysis (HD) in the United States.[1] Symptom burden is high in HD patients and patients on average report a median of nine symptoms over a week.[2] The Standardized Outcomes in Nephrology (SONG)-HD, has identified outcomes that are important to physicians and patients.[3] While cardiovascular disease and mortality outcomes are easily tracked and identified, symptoms are difficult to identify and usually require prospective survey of patients or manual chart review which are time consuming (many surveys being over 30 questions), and only provides a cross sectional view.[4,5]

Electronic health records (EHRs) have been widely implemented in most hospital systems and dialysis units.[6] At each HD session, patients are regularly observed for adverse signs and symptoms by nurses, technicians, and physicians. These encounters are documented in EHRs as “free text” and infrequently as structured data.[7] Natural language processing (NLP) allows for the ‘reading’ of unstructured documentation and converts it into discrete data for analysis. We sought to determine the ability of NLP to identify fatigue, nausea and/or vomiting(N/V), anxiety, depression, itching, cramps, and pain from the EHR of HD patients. We then compared the performance of NLP and ICD against manual chart review.

Results:

Patient Characteristics:

We identified 1080 patients receiving maintenance HD from Bio*Me* (Figure S1 A). 46 of these patients who enrolled after 2017 served as a separate validation dataset. Patients had a

mean age of 64±13 years, 42% were women, and 42% self-reported as African American. There was a high prevalence of diabetes (65%), hypertension (88%), coronary artery disease (40%), and congestive heart failure (32%) (Table 1). The median number of encounters was 109, (interquartile range (IQR) 41–241), progress notes were 342 (IQR 102–782) and discharge summaries were 16 (IQR 2–54). The mean follow up time was 8.7±5.5 years (Table 1). From MIMIC-III, we identified 519 chronic HD patients utilizing ICD-9 codes (Figure S1 B). The mean age of patients was 70 ± 39.6 years, 41% of patients were women, and 63% self-reported as white. Prevalence of co-morbidities were high, diabetes (54%), hypertension (91%), coronary artery disease (46%), and congestive heart failure (47%) (Table 1). Median progress note count was 10 (IQR 0–55) and median discharge summary count was 1 (IQR 1–2). As a majority of patients only had 1 encounter, follow up time could not be calculated.

Symptom Identification using NLP vs Administrative Codes:

In the Bio*Me* development cohort, NLP identified symptoms more frequently than did ICD codes (Figure 1 A). The most frequent symptoms identified were pain (NLP 93% vs. ICD 46%, P<0.001), fatigue (NLP 84% vs. ICD 41%, P<0.001), and N/V (NLP 74% vs. ICD 19%, P<0.001). Symptoms were most frequently identified from progress notes (39%−84%) and discharge summaries (14%−33%). When normalized by number of encounters and follow up time in the Bio*Me* development cohort, the mean frequency of symptoms were 0.8, 0.5, 0.5, 0.4, 0.1, 0.07, and 0.003 encounters/year for pain, fatigue, depression, itching, anxiety, N/V, and for cramping, respectively. In the Bio*Me* validation cohort, the mean frequency of symptoms were 0.1, 0.02, 0.01, 0.01, 0.006, 0.003, 0.001 encounter/year for pain, depression, fatigue, anxiety, N/V, itching, and cramping, respectively.

In MIMIC-III, the most common symptoms by NLP were pain (NLP 94% vs. ICD 6%, P=0.15), fatigue (NLP 62% vs. ICD 1%, P=0.05), and N/V(NLP 56% vs. 3%, P =0.003) (Figure 1 B).. Depression, anxiety, and pain were the most common symptoms by ICD codes (all 6%).

Manual Chart Validation of 50 Randomly Selected Charts:

In the Bio*Me* development cohort, agreement across investigators for chart review was high (kappa statistic 0.6–1). Frequency of symptoms identified by NLP+ICD+manual review was 4%−54%, for NLP+manual review was 16%−54%, and ICD+manual review was 0–2% (Figure 2). Sensitivity for NLP ranged from 0.85 (95% CI 0.65–96) for depression to 0.99 (95% CI 0.93–1) for fatigue while sensitivity for ICD ranged from 0.09 (95% CI 0.01–0.29) for cramps to 0.59 (95% CI 0.43–0.73) for fatigue. Specificity for NLP ranged from 0.5 (95% CI 0–1) for pain to 0.96 (95% CI 0.8–1) for itching, while specificity for ICD ranged from 0.5 (95% CI 0.37–0.66) for pain to 0.98 (95% CI 0.86–1) for itching (Figure 3 A and Table S1 A). ICD codes were significantly more specific for N/V(NLP 0.57 (95% CI 0.29–0.82) vs. ICD 0.97 (95% CI 0.77–1), P=0.03). F1 scores for NLP ranged from 0.82 to 0.99 and were significantly higher than ICD for all symptoms (0.28 – 0.83). The addition of medications to ICD codes for identification of N/V, anxiety, depression and pain improved sensitivity of ICD alone however worsened specificity (Figure S2 A).

In the Bio*Me* validation cohort, the sensitivity for NLP ranged from 0.78 (95% CI 0.52–0.94) for depression to 0.99 (95% CI 0.92–1) for fatigue while sensitivity of ICD ranged from 0.13 (95% CI 0.02–0.27) for cramp to 0.71 (95% CI 0.56–0.85) for fatigue. (Table S1 B).

Twenty-five patients were identified to have undergone PHQ-9 depression screening, of which 24 patients were identified to have depression PHQ-9 and/or clinical history. NLP correctly identified 22 (92%) patients while ICD 9/10 identified 20 (83%) patients. In MIMIC-III, sensitivity for NLP ranged from 0.5 for cramp to 0.98 (95% CI 0.86–1) for fatigue while sensitivity for ICD ranged from 0.04 (95% CI, 0–0.21) for fatigue to 0.5 (95% CI N/A) for cramp (Figure 2 B and Table S1 C). Specificity for NLP ranged from 0.11 (95% CI 0–0.48) for pain to 0.98 (95% CI 0.89–1) for itching, while for ICD it was 0.95 (95% CI 0.66–1) for pain to 0.99 (95% CI 0.93–1) for itching. ICD had significantly higher specificity for fatigue (NLP 0.77 (95% CI 0.56–0.91) vs ICD 0.98 (95% CI 0.87–1), P=0.03), depression (NLP 0.81 (95% CI 0.64–0.92) vs. ICD 0.99 (95% CI 0.9–1), P=0.02), and pain (NLP 0.11 (95% CI 0–0.48) vs. ICD 0.95 (95% CI 0.66–1), P<0.001) in MIMIC-III.

Symptom Burden

In Bio*Me*, NLP identified 44 (4%) patients, 61 (6%) patients, 87 (8%) patients, 99 (10%) patients, 177 (17%) patients, 158 (15%) patients, 204 (20%) patients, and 204 (20%) patients with 0,1,2,3,4,5,6, and 7 symptoms respectively. Patients who did not have any symptoms identified by NLP had a median of 7 (IQR 2–40) encounters/year, while patients with all 7 symptoms had a median of 24 (IQR 15–43) encounters/year. There was a moderate correlation between the number of encounters/year and the number symptoms identified by NLP (correlation coefficient 0.36, P<0.001). Within the 50 patients that had manual chart review, symptom burden identified by NLP and manual chart review was similar, however patients had less symptoms by ICD codes than manual chart review (Figure 4 A, B, C). There was a moderate significant positive correlation between number of encounters per year and number of symptoms identified by NLP (correlation coefficient 0.5, P<0.001), ICD (correlation coefficient 0.35, P=0.01), and manual chart review (correlation coefficient 0.5, P<0.001) in Bio*Me*. In MIMIC-III, NLP identified 21 (4%) patients, 51 (10%) patients, 136 (26%) patients, 122 (24%) patients, 98 (19%) patients, 65 (13%) patients, 21 (4%) patients, and 5 (1%) patients with 0,1,2,3,4,5,6, and 7 symptoms respectively.

Subgroup Analysis in BioMe:

608 participants had at least 2 years of follow up time. When restricted to only 1 year of notes, NLP identified symptoms less frequently than without date restrictions (fatigue 58% vs. 84%, N/V39% vs. 74%, anxiety 26% vs. 54%, depression 18% vs. 55%, itching 22% vs. 48%, and cramp 16% vs. 44%) except for pain which was found at a similar rate (90% in 1 year subset vs. 93% no restrictions) (Figure S3). Even with date restrictions, NLP identified more symptoms and had better sensitivity than ICD codes (Table S1 D).

1024 participants had at least 1 encounter with notes available, these patients were then grouped into tertiles (low (≤62 encounters), medium (63–186 encounters), and high (≥187 encounters)) based on the number of encounters. An increasing number of symptoms were identified using NLP and ICD with increasing encounters for all symptoms except for pain. There was no substantial increase in identification of pain with NLP between medium and high encounter groups. Regardless of symptom and encounter group, NLP identified more symptoms than ICD did (Figure S4).

Out of 100,118 notes, 11,066 (11%) of notes were from an inpatient hospital stay. Symptoms were identified more frequently in outpatient notes than inpatient notes except for N/V. Fatigue identification had the largest difference by inpatient and outpatient notes, with a difference of 19% (Figure S5). Overall, NLP had better sensitivity for symptom identification in outpatient notes but better specificity in inpatient notes (Figure S2 B).

A total of 533 (52%) patients had 7,476 episodes of pain with severity documented; 3,137 (42%) were mild, 2,232 (30%) were moderate, and 2,107 (28%) were severe. NLP performed similarly across pain severity types (Figure S2 C).

Discussion:

Although symptoms are common in HD patients and are identified as important to patients and providers, efficient retrospective assessment of symptoms from EHR is difficult. We show that NLP has better sensitivity than ICD codes at identifying seven common symptoms in patients from the Bio*Me* Biobank with validation of results in a separate validation cohort from Bio*Me* and an external cohort from MIMIC-III.[3] The symptom burden was high, with a majority of patients having at least 4 or more symptoms identified by NLP. Finally, there was a positive correlation between number of encounters and number of symptoms identified by NLP.

The SONG-HD initiative identified several outcomes important to all stakeholders and has emphasized the importance of clinical research that includes these symptoms. Prior research that employed patient-centered outcomes as endpoints have required prospective surveys for their execution.[2,8] Alternatives to this approach include chart validation potentially with aid of computer text searching and chart review tools. However, these methods are labor and time intensive. While NLP can process notes in an efficient manner, there are few studies in nephrology that have utilized NLP and included symptoms or patient-centered outcomes.[9–14]

Symptom prevalence identified in the Bio*Me* cohort by NLP is similar to prior published survey data on symptoms.[2,8,16] Symptoms such as itching and cramps were less frequent, while other symptoms such as N/V were found more commonly. Differences in patients enrolled in studies and those seen in real world practice likely contributes to the differences in prevalence. Additionally, the ethnically and racially diverse nature of the Bio*Me* Biobank and the critically ill nature of MIMIC-III patients are likely contributors to differences in symptom prevalence. How the prevalence of symptoms identified here compares to the general outpatient U.S. HD population needs to be further elucidated.

While surveys done in prior studies have been in patients who are stable at their outpatient hemodialysis centers, we included outpatient and acute in-patient notes. On subgroup analysis, we found that more symptoms were picked up in the outpatient notes than the inpatient notes. The larger proportion of notes for outpatient encounters likely contributes to the higher sensitivity in outpatient notes. Additionally providers are more likely to discuss overall health in the outpatient setting when the patient is not acutely ill, while in the inpatient notes providers will focus on the admitting diagnosis. As MIMIC-III consists of progress notes from critically ill patients, symptoms were identified at an even lower rate. This is likely related to the patients being critically ill and unable to verbalize their symptoms along with the providers focus on admission diagnosis and contributing comorbidities instead of symptoms and psychosocial comorbidities.

We chose not to place limitations on the number, timing, or type of notes, which may have increased the likelihood of NLP or ICD codes identifying a symptom. However, in sensitivity analysis, NLP consistently identified more symptoms than ICD codes. While the false positives may be contributing to this difference, we suspect this is a small contributor given relatively small differences in specificity between NLP and ICD. Additionally, the lack of note restrictions may lead to identification of symptoms that are not caused by patient's ESRD status. However, these symptoms remain important patient outcomes as they were deemed to be important to patients, physicians, and caregivers.[3]

We found that NLP out performed ICD codes for symptom identification.[17–20] As ICD codes are administrative and billing codes clinicians may be less inclined to use them to document symptoms experienced by HD patients, especially if they do not count towards overall reimbursement. Sensitivity for ICD codes are generally moderate even in more common conditions such as myocardial infarction (72%) and hypertension (78%).[21] The addition of medications to ICD codes for identification of N/V, anxiety, depression, and pain increased sensitivity but decreased specificity likely due to the use of medications for other indications (e.g. bupropion for depression and also for smoking cessation).

ICD codes for symptoms had high specificity and high PPV. Therefore, the NLP method may be favored for identification of a large cohort of patients with symptoms accepting the risk of higher false positives while the ICD method may be favored for identification of patients highly likely to have symptoms accepting the higher false negative rate. While NLP was more sensitive at identifying depression, ICD codes were more specific. This was due to false positives for depression used in other clinical contexts (e.g. depressions on electrocardiogram or temporal depressions). While we could potentially improve the specificity of NLP for depression by excluding specific phrases found during chart review this is likely to reduce the generalizability of NLP for external cohorts due to variability in provider documentation across institutions.

Our study should be interpreted in the light of some limitations including the dependence of symptom identification on the number of encounters and notes available. However, this is a common issue with EHR systems, where both sicker patients as well as patients with longer length of follow up have more data.[22] Unfortunately, data regarding the author of the note is not available and we cannot comment on the documentation of symptoms by provider type.

Additionally, only symptoms which the provider is screening for are documented and therefore NLP may miss those symptoms patients are not discussing with their providers which may lead to an underestimation of symptom prevalence.[23] Neither the Bio*Me* nor MIMIC-III datasets are exclusive to outpatient HD patients, which make comparison with prior published data difficult and reduce the generalizability of our results. However, the prevalence of symptoms in our study is similar to prior published survey data.[2,16] Data is extracted from EHR of respective institutions, and this export process may affect the generalizability of results to other institutions. While presence of symptoms varied throughout time, we chose to classify patients as ever present or never present as our goal was to evaluate the performance of NLP for identifying patients with the symptoms. In our subgroup analyses of one year of notes, we looked only at the date of the note and not whether the query was flagged as a current or past temporal context, which may change the frequency of symptoms identified. Unfortunately, as we did not have concurrent survey data available, we used manual chart review as our gold standard which may be imperfect., The results of our test statistics were relatively consistent across Bio*Me* and MIMIC-III cohorts, suggesting that our NLP algorithm could have generalizability across different medical systems. Unfortunately, we did not perform formal error analysis but further work on NLP methods could benefit from formal error analysis.

In conclusion, we utilized NLP to identify important patient symptoms from EHR of HD patients from the Mount Sinai health system and validated our results in MIMIC-III. NLP out performed ICD codes for identification in regards to sensitivity, negative predictive value (NPV), and F1 score for a majority of symptoms in both the cohorts. Additional refinement of NLP approaches and testing in the EHR of outpatient HD units is needed to further validate our findings and to utilize NLP approaches in the care of our patients.

Methods:

Study Population

From a cohort of 38,575 participants from the Bio*Me* Biobank at Mount Sinai, we retrieved all notes of Bio*Me* participants available from a centralized data mart from January 1, 2010 up to March 15, 2019. The Bio*Me* Biobank is a prospective registry of patients from the Mount Sinai Healthcare System linked to the United States Renal Data System (USRDS). We included patients on HD excluding those with a kidney transplant and never on dialysis. As linkage information did not include dialysis type or dialysis access type, peritoneal (PD) patients were excluded using ICD codes as PD. (Table S2 A). The NLP development cohort included only patients who enrolled in Bio*Me* prior to December 31,2017. The institutional review board approved the Bio*Me* protocols and informed consent was obtained for all subjects.

We validated performance of our NLP algorithm using two distinct cohorts (1) Bio*Me* validation cohort comprised of chronic HD patients from Bio*Me* that were not included in the original development cohort and (2) the Medical Information Mart for Intensive Care (MIMIC-III) database.[24] The Bio*Me* validation cohort were patients who enrolled in Bio*Me* between January 1, 2018 and March 15, 2019 and were identified using ICD 9/10 codes (Table S2 A). MIMIC-III is a critical care database of patients from a large, single center

tertiary care hospital from 2001–12.[24] We included all notes from the MIMIC–III database. Since MIMIC-III could not be linked with USRDS, ESKD was identified as patients who had an ESKD code and a code for dialysis procedure or diagnosis after excluding patients with acute kidney injury codes and PD patients by PD procedure codes (Table S2 A). As MIMIC-III is a de-identified publically available database, evaluation of data from this source was considered IRB exempt.

Patient comorbidities were identified utilizing the Clinical Classification Software (CCS) developed by the Healthcare Cost and Utilization Project.[25] The CCS aggregates ICD codes into clinically meaningful and mutually exclusive categories. The codes used for identification are included in Table S2 B.

Study Design

This is a retrospective cohort study of HD patients drawn from the EHR from two medical systems. We utilized the CLiX NLP engine produced by Clinithink (London, UK) to parse notes. CLiX NLP is a NLP software that matches free text to Systematized Nomenclature of Medicine - - Clinical Terms (SNOMED CT).[26] SNOMED CT is a comprehensive healthcare terminology resource that has an inherent hierarchy consisting of overarching concepts, i.e. parent terms, which encompass more specific concepts, i.e., children terms. Figure S6 includes an example of how “cramp” would be represented in the SNOMED CT hierarchy. In our testing for this and other projects we found that CLiX NLP was able to handle typographical errors, sentence context, and negation well.[27,28] Common abbreviations (e.g. N/V for nausea and/or vomiting) were correctly identified; however during chart review additional abbreviations that were incorrectly identified required a request to alter the NLP algorithm. CLiX NLP identifies terms such as “no”, “denies”, “not” as negative and applies it to the SNOMED CT thereby marking the query as “present” or “absent”. Therefore, we did not use specific negative terms for exclusion, instead only those marked as “present” were considered as positive. There was no restriction on number of notes or types of notes placed. Note types included progress notes (from all providers including social worker, physical therapy, nursing, and physician), radiology reports, discharge summaries, and pathology reports.

We queried for fatigue, depression, pain, N/V, anxiety, and cramps[3]. SNOMED CT for the associated outcomes were selected through extensive review by two physicians (Table S3). These specific terms were selected due to their inability to be identified from structured data.

We used a SNOMED CT query engine (a second component of CLiX) to perform hierarchical subsumption queries to identify all relevant SNOMED CT, both parent terms and the associated children terms for each outcome. This was first identified on the document level and then on the patient level. For depression, a chronic disease, NLP identification on at least two different dates was necessary to be considered positive; for all other symptoms identification on one note was considered positive. CLiX NLP reads through each sentence to identify all associated SNOMED CT. Then CLiX NLP’s inherent description logic outputs details associated with each term including subject, temporality, present vs. absence, and if appropriate location/laterality (Figure S7). A query was considered positive if the subject was identified as the subject of record and it had a known

present context. Date of query positive was date of the note. While CLiX NLP is a proprietary system, this study can be replicated with other NLP tools that utilize SNOMED CT. We chose this NLP method due to the research teams' familiarity with it and availability to us. Other valid methods, including machine learning, were not used due to lack of mature methodologies for this specific project.

We performed two iterations of NLP parsing with manual chart review of 50 randomly selected charts, a test set, guiding the second iteration. We rectified errors in identification in the NLP engine prior to the execution of the final parsing. Examples included phrases such as "The patient was advised to call for any fever or for prolonged or severe pain or bleeding" and "EKG sinus tach with V4, V5 depressions". We modified the NLP algorithm to recognize these as negative expressions. We report results in this manuscript from the final NLP query with test statistics calculated from a separate manual chart review of 50 randomly selected charts as described below. Examples of false positives that were identified during this final chart review are presented in Table S4. This final NLP algorithm was then validated in 46 distinct HD patients from Bio*Me* and MIMIC-III.

We compared performance of ICD-CM codes with the results obtained from CLiX NLP.[29–31] ICD-9 and 10 codes were used in Bio*Me* while only ICD-9 codes were available in MIMIC-III (Table S2 C). To determine if medication data improved ICD identification of symptoms, we identified medications used for pain, N/V, anxiety, and depression from RxNorm (Table S5) and identified patients who were ever prescribed these medications.[32] Medications which are commonly used for other indications (e.g. aspirin for 2nd prevention of cardiac events) was removed from the list. Finally, both methods were compared with independent chart review by two physicians. We randomly selected 50 patient charts from Bio*Me* and MIMIC-III, using SAS (PROC SURVEY SELECT method SRS) to perform simple random sampling. Then all notes from the same 50 charts were reviewed for all symptoms. All patients from the Bio*Me* validation cohort underwent manual chart review. When there was disagreement between manual validations for a patient, joint review of the patient's chart was performed until consensus agreement was obtained.

To evaluate NLP performance across note types, notes were categorized into inpatient or outpatient. Manual chart review of 50 randomly selected charts was performed. Next, we looked at symptom identification within progress notes, discharge summaries, pathology reports and radiology reports. For pain, we extracted severity and categorized it into mild, moderate, and severe. Manual chart review of 25 randomly selected cases for each severity and 25 randomly controls (those with pain but no severity identified) was performed.

Two additional subgroup analyses were performed using data from Bio*Me* patients. First, we restricted NLP to only one year of notes from patients who had at least two years of data. Manual chart review was done on 30 patients. Second, only patients with at least 1 encounter with notes available were included and grouped into tertiles basded on the number of encounters (low, medium, and high). Unfortunately, the MIMIC-III database was solely an ICU database and therefore lacks the repeated encounters and longitudinal follow up that is available in Bio*Me*, therefore these subgroup analyses could not be performed.

Lastly, we compared NLP depression positive with Patient Health Questionnaire 9 (PHQ-9) screening documentation[33,34] We considered depression screening positive if patients scored ≥10 or there was evidence of history of depression (i.e. cognitive behavior therapy, anti-depressive medications, or prior suicide attempts).

Statistical Analysis:

We calculated sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 scores of NLP and ICD9/10 codes. F1 scores were calculated as a measure of accuracy that considers both the sensitivity and PPV.[35] For cells on the 2×2 table where the value was 0, we adapted the Woolf-Haldane correction method for logistic regression and entered 0.5 to allow for calculation of test statistics.[36,37] 95% CI were calculated using the PROC FREQ procedure in SAS using the binomial option.[38] We compared estimates of sensitivity, specificity using the McNemar's test with significance set using a two sided p value of <0.05. We compared NPV and PPV using the generalized score statistic method and the SAS macro created by Gondara et al.[39] Unfortunately 95% CI and P values could not be generated if 2 or more cells in the 2×2 table were empty. We calculated Pearson correlation coefficient to determine the correlation between number of encounters and number of symptoms identified by NLP. We performed all analysis using SAS version 9.4 (SAS Institute, Cary NC) and R 3.6.0 (R Foundation for Statistical Computing, Vienna, Austria).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We want to thank all participants of Bio*Me* and MIMIC-III.

Sources of support: LC is supported in part by the NIH (5T32DK007757 - 18). G.N.N. is supported by a career development award from the National Institutes of Health (NIH) (K23DK107908) and is also supported by R01DK108803, U01HG007278, U01HG009610, and U01DK116100. S.G.C. is supported by the following grants from the NIH: U01DK106962, R01DK106085, R01HL85757, R01DK112258, and U01OH011326.

Disclosures:

L.C. is supported in part by the NIH (5T32DK007757 - 18). G.N.N. and S.G.C. are co-founders of RenalytixAI and G.N.N. and S.G.C. are members of the advisory board of RenalytixAI and own equity in the same. G.N.N has received operational funding from Goldfinch Bio. G.N.N has received consulting fees for BioVie Inc. S.G.C. has received consulting fees from Goldfinch Bio, CHF Solutions, Quark Biopharma, Janssen Pharmaceuticals, and Takeda Pharmaceuticals. G.N.N. and S.G.C. are on the advisory board for pulseData and have received consulting fees and equity in return. G.N.N. is supported by a career development award from the National Institutes of Health (NIH) (K23DK107908) and is also supported by R01DK108803, U01HG007278, U01HG009610, and U01DK116100. S.G.C. is supported by the following grants from the NIH: U01DK106962, R01DK106085, R01HL85757, R01DK112258, and U01OH011326. T.V.V. was part of launching Clinithink and retains a financial interest in the company.

References

1. United States Renal Data System. 2018 USRDS annual data report: Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2018.

2. Weisbord SD, Fried LF, Arnold RM et al. Prevalence, Severity, and Importance of Physical and Emotional Symptoms in Chronic Hemodialysis Patients. J. Am. Soc. Nephrol 2005; 16: 2487–2494. [PubMed: 15975996]

3. Tong A, Manns B, Hemmelgarn B et al. Establishing Core Outcome Domains in Hemodialysis: Report of the Standardized Outcomes in Nephrology–Hemodialysis (SONG-HD) Consensus Workshop. Am. J. Kidney Dis. 2017; 69: 97–107. [PubMed: 27497527]

4. Weisbord SD, Fried LF, Arnold RM et al. Development of a symptom assessment instrument for chronic hemodialysis patients: the dialysis symptom index. J. Pain Symptom Manage. 2004; 27: 226–240. [PubMed: 15010101]

5. Hays RD, Kallich JD, Mapes DL et al. Development of the kidney disease quality of life (KDQOL) instrument. Qual. Life Res 1994; 3: 329–38. [PubMed: 7841967]

6. Adler-Milstein J, DesRoches CM, Furukawa MF et al. More than half of US hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. Health Aff. (Millwood) 2014; 33: 1664–71. [PubMed: 25104826]

7. Hernandez-Boussard T, Tamang S, Blayney D et al. New Paradigms for Patient-Centered Outcomes Research in Electronic Medical Records: An Example of Detecting Urinary Incontinence Following Prostatectomy. EGEMS (Washington, DC) 2016; 4: 1231.

8. Merkus MP, Jager KJ, Dekker FW et al. Nephrology Dialysis Transplantation Physical symptoms and quality of life in patients on chronic dialysis : results of The Netherlands Cooperative Study on Adequacy of Dialysis (NECOSAD). Nephrol. Dial. Transplant 1999: 1163–1170. [PubMed: 10344356]

9. Perotte A, Ranganath R, Hirsch JS et al. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J. Am. Med. Inform. Assoc 2015; 22: 872–80. [PubMed: 25896647]

10. Singh K, Betensky RA, Wright A et al. A Concept-Wide Association Study of Clinical Notes to Discover New Predictors of Kidney Failure. Clin. J. Am. Soc. Nephrol 2016; 11: 2150–2158. [PubMed: 27927892]

11. Chase HS, Radhakrishnan J, Shirazian S et al. Under-documentation of chronic kidney disease in the electronic health record in outpatients. J. Am. Med. Inform. Assoc 2010; 17: 588–94. [PubMed: 20819869]

12. Nadkarni GN, Gottesman O, Linneman JG et al. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. AMIA Annu. Symp. Proc 2014; 2014: 907–16. [PubMed: 25954398]

13. Malas MS, Wish J, Moorthi R et al. A comparison between physicians and computer algorithms for form CMS-2728 data reporting. Hemodial. Int 2017; 21: 117–124. [PubMed: 27353890]

14. Nigwekar SU, Solid CA, Ankers E et al. Quantifying a Rare Disease in Administrative Data: The Example of Calciphylaxis. J. Gen. Intern. Med 2014; 29: 724–731.

15. Hernandez-Boussard T, Kourdis PD, Seto T et al. Mining Electronic Health Records to Extract Patient-Centered Outcomes Following Prostate Cancer Treatment. AMIA … Annu. Symp. proceedings. AMIA Symp 2017; 2017: 876–882.

16. Caplin B, Kumar S, Davenport A. Patients' perspective of haemodialysis-associated symptoms. Nephrol. Dial. Transplant 2011; 26: 2656–2663. [PubMed: 21212166]

17. Waikar SS, Wald R, Chertow GM et al. Validity of International Classification of Diseases, Ninth Revision, Clinical Modification Codes for Acute Renal Failure. J. Am. Soc. Nephrol 2006; 17: 1688–94. [PubMed: 16641149]

18. Vlasschaert MEO, Bejaimal SAD, Hackam DG et al. Validity of administrative database coding for kidney disease: A systematic review. Am. J. Kidney Dis 2011; 57: 29–43. [PubMed: 21184918]

19. Semins MJ, Trock BJ, Matlaga BR. Validity of administrative coding in identifying patients with upper urinary tract calculi. J. Urol 2010; 184: 190–2. [PubMed: 20478584]

20. McCormick N, Lacaille D, Bhole V et al. Validity of Heart Failure Diagnoses in Administrative Databases: A Systematic Review and Meta-Analysis. Guo Y, ed. PLoS One 2014; 9: e104519. [PubMed: 25126761]

21. Quan H, Sundararajan V, Halfon P et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data.; 2005.

22. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA … Annu. Symp. proceedings. AMIA Symp 2013; 2013: 1472–7.
23. Weisbord SD, Fried LF, Mor MK et al. Renal provider recognition of symptoms in patients on maintenance hemodialysis. Clin. J. Am. Soc. Nephrol 2007; 2: 960–7. [PubMed: 17702730]
24. Johnson AEW, Pollard TJ, Shen L et al. MIMIC-III, a freely accessible critical care database. Sci. Data 2016; 3: 160035. [PubMed: 27219127]
25. HCUP CCS. Healthcare Cost and Utilization Project (HCUP) 5 2016 Agency for Healthcare Research and Quality, Rockville, MD www.hcupus.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed October 3, 2016.
26. Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. Proc. a Conf. Am. Med. Informatics Assoc. AMIA Fall Symp 1997: 640–4.
27. Van Vleck TT, Chan L, Coca SG et al. Augmented intelligence with natural language processing applied to electronic health records for identifying patients with non-alcoholic fatty liver disease at risk for disease progression. Int. J. Med. Inform 2019; 129: 334–341. [PubMed: 31445275]
28. Clark MM, Hildreth A, Batalov S et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci. Transl. Med 2019; 11: eaat6177. [PubMed: 31019026]
29. Fiest KM, Jette N, Quan H et al. Systematic review and assessment of validated case definitions for depression in administrative data. BMC Psychiatry 2014; 14: 289. [PubMed: 25322690]
30. Tian TY, Zlateva I, Anderson DR. Using electronic health records data to identify patients with chronic pain in a primary care setting. J. Am. Med. Informatics Assoc 2013; 20: e275–e280.
31. Kisely S, Lin E, Gilbert C et al. Use of Administrative Data for the Surveillance of Mood and Anxiety Disorders. Aust. New Zeal. J. Psychiatry 2009; 43: 1118–1125.
32. RxClass.
33. Kroenke K, Spitzer RL. The PHQ-9: A New Depression Diagnostic and Severity Measure. Psychiatr. Ann 2002; 32: 509–515.
34. Watnick S, Wang P-L, Demadura T et al. Validation of 2 Depression Screening Tools in Dialysis Patients. Am. J. Kidney Dis 2005; 46: 919–924. [PubMed: 16253733]
35. Sasaki Y, Fellow R. The truth of the F-measure.; 2007.
36. Lawson R Small Sample Confidence Intervals for the Odds Ratio. Commun. Stat. - Simul. Comput 2004; 33: 1095–1113.
37. Dureh N, Choonpradub C, Tongkumchum P. An alternative method for logistic regression on contingency tables with zero cell counts.
38. 24170 - Estimating sensitivity, specificity, positive and negative predictive values, and other statistics.
39. Gondara L A SAS ® macro to compare predictive values of diagnostic tests.

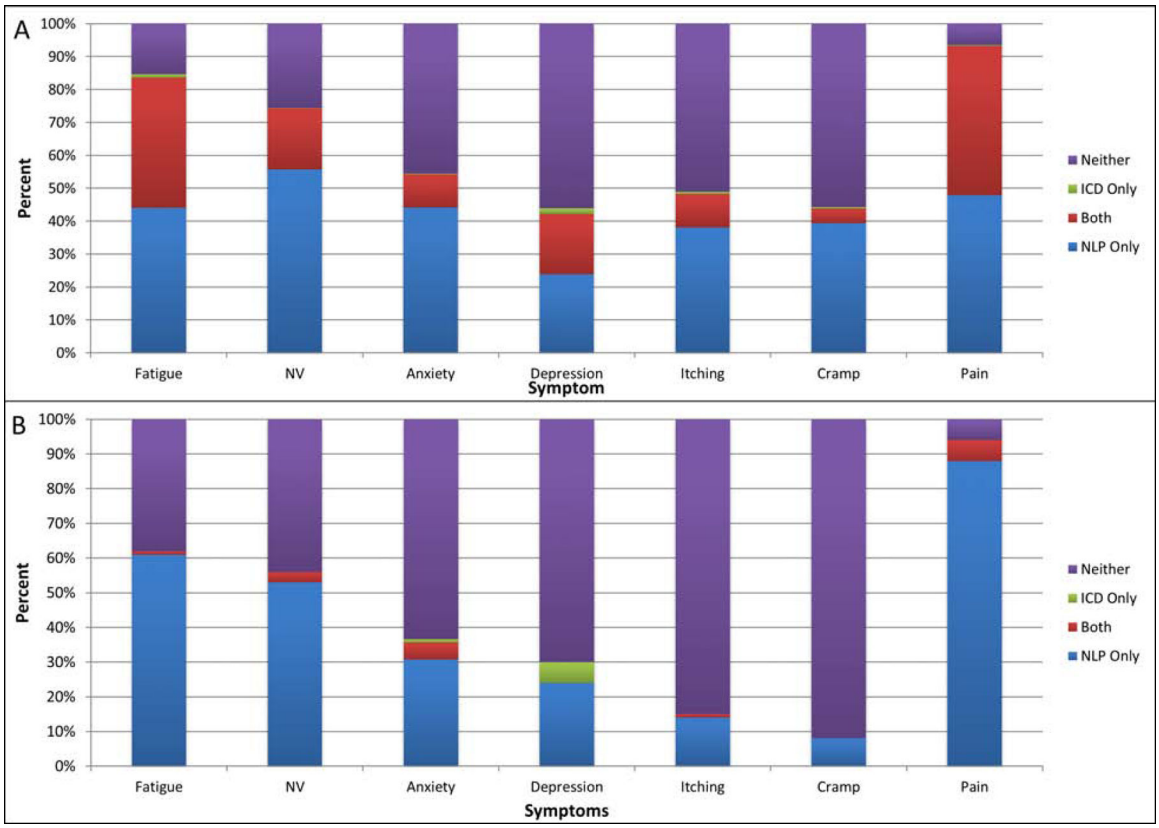


Figure 1:

Frequency of symptom identified by NLP and ICD from (A) Bio*Me* and (B) MIMIC-III. Blue bar indicates percentage of patients where symptom was found only by NLP, green bar indicates percentage of patients where symptom was found by only by ICD, red bar indicates percentage of patients where symptom was found by both NLP and ICD, while purple bar indicates percentage of patients where the symptom was found by neither NLP or ICD

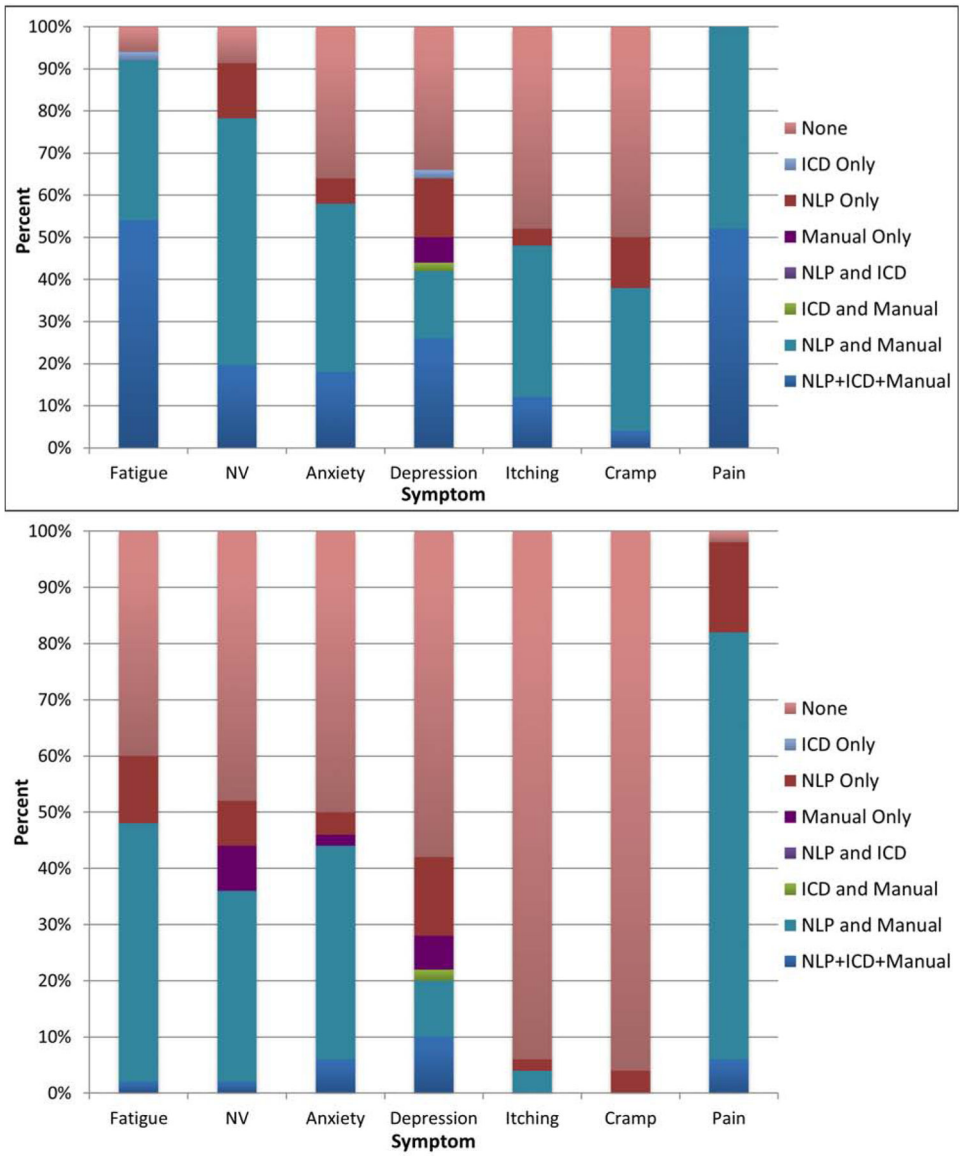


Figure 2:

Frequency of symptoms from 50 patients from A) BioMe and B) MIMIC-III who had manual chart review.

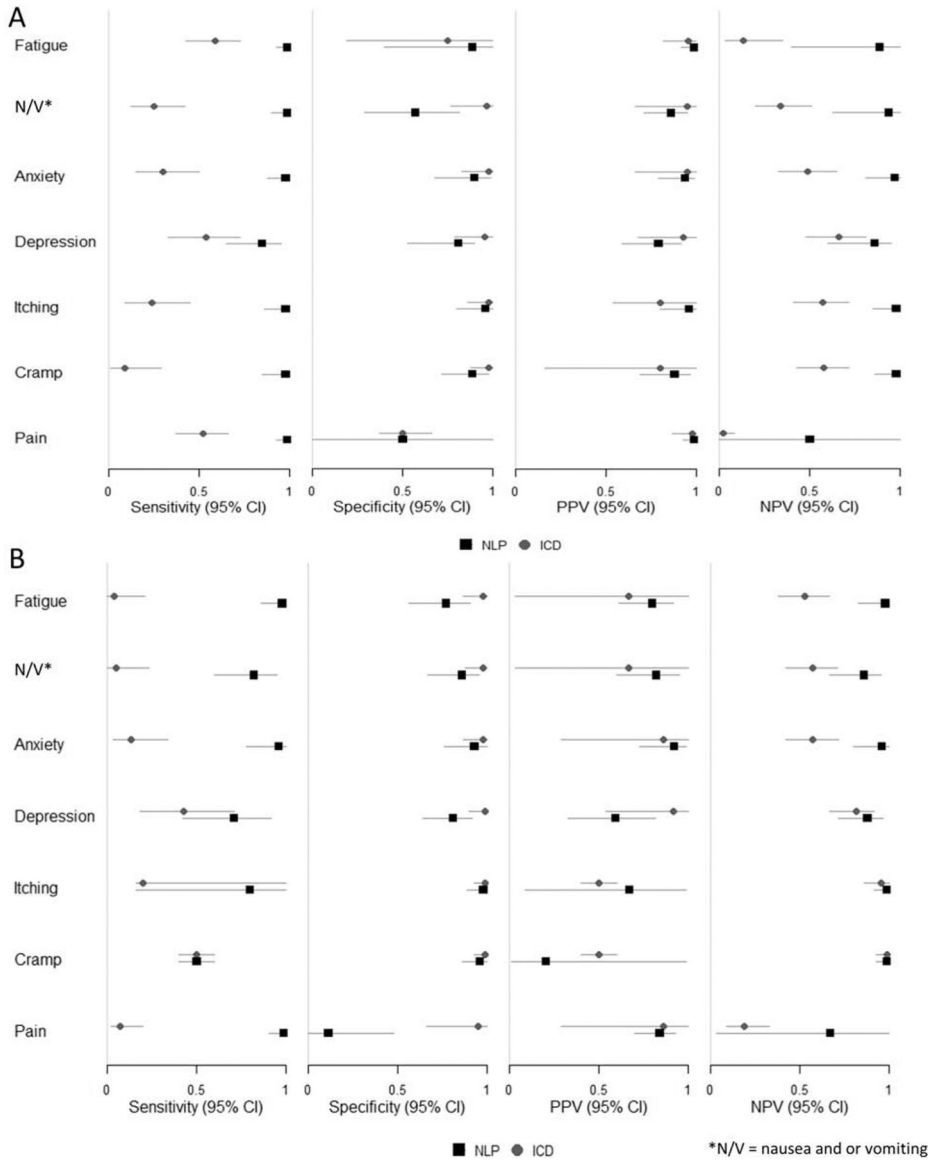


Figure 3:

Sensitivity, specificity, PPV, NPV, and F1 score of NLP vs. ICD for identification of symptoms for A) BioMe and B) MIMIC-III calculated using manual chart review of 50 patient charts

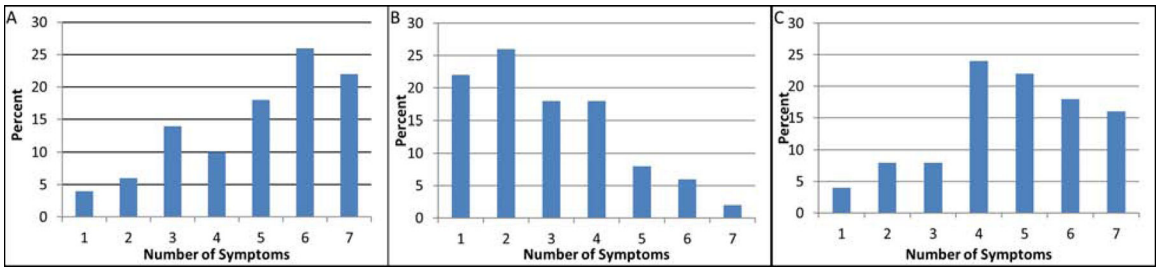


Figure 4:

Overall symptom burden demonstrating the number of symptoms identified from 50 BioMe patients by A) NLP, B) ICD, and C) manual review

Table 1:

Patient Characteristics of Bio*Me* and MIMIC-III

	Bio*Me* Development (n=1034)	**MIMIC-III (n=519)**	**Bio*Me* Validation (n=46)**
Age [years]	64 ±13.3	70±39.6	57±12.6
Female	433 (42)	212 (41)	18 (39)
Race/ethnicity:			
African American	433(42)	97 (19)	17 (40)
European American	146(14)	329 (63)	4 (9)
East Asian	14(1.4)	18 (3)	1 (2)
Hispanic	376(36)	26 (5)	21 (46)
Missing	2(0.2)	20 (4)	0 (0)
Other	63(6)	29 (6)	3 (7)
Comorbidities:			
Diabetes	671(65)	278 (54)	29 (63)
Hypertension	915(88)	473 (91)	44 (96)
Coronary artery disease	412(40)	241 (46)	14 (30)
Congestive heart failure	334(32)	243 (47)	13 (28)
Insurance Type:			
Medicare	461 (45)	387 (75)	17 (40)
Medicaid	332 (32)	39 (8)	20 (46)
Private	215 (21)	84 (16)	5 (12)
Other/missing	26 (3)	9 (2)	1(9)
Note types [median (IQR)]:			
Progress Notes	342 (102–782)	10 (0–55)	366 (234–486)
Discharge Summaries	16 (2–54)	1 (1–2)	1 (1–1)
Radiology	48 (14–105)	13 (4–33)	14.5 (6–28)
Pathology/Test Report	0 (0–1)	9 (4–20)	1 (1–3)
Mean follow up time [years]	8.7±5.5	N/A *	6.9±3.5

Data are shown as mean ± standard deviation or count (%) except where specified

* As a majority of patients only had 1 encounter, follow up time was not calculated.