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

see commentary on page 263

Lili Chan[1,2], Kelly Beers[1], Amy A. Yau[1], Kinsuk Chauhan[1], Áine Duffy[2], Kumardeep Chaudhary[2], Neha Debnath[1], Aparna Saha[2], Pattharawin Pattharanitima[1], Judy Cho[2], Peter Kotanko[1,3], Alex Federman[4], Steven G. Coca[1], Tielman Van Vleck[2,5] and Girish N. Nadkarni[1,2,5]

[1]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [2]The Charles Bronfman Institute for Personalized Medicine, Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [3]Renal Research Institute, New York, New York, USA; and [4]Division of General Internal Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA

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 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.

Kidney International (2020) **97,** 383–392; https://doi.org/10.1016/j.kint.2019.10.023

Correspondence: *Lili Chan or Girish N. Nadkarni, Icahn School of Medicine at Mount Sinai, One Gustave L Levy Place, Box 1243, New York, New York 10029, USA. E-mail: lili.chan@mountsinai.org or girish.nadkarni@mountsinai.org*

[5]TVV and GNN contributed equally.

Received 30 April 2019; revised 27 September 2019; accepted 18 October 2019; published online 9 November 2019

KEYWORDS: geriatric nephrology; hemodialysis; natural language processing; patient-centered outcomes; symptoms

Copyright © 2019, International Society of Nephrology. Published by Elsevier Inc. All rights reserved.

More than 450,000 patients are undergoing maintenance hemodialysis (HD) in the United States.[1] Symptom burden is high in patients undergoing HD, and patients on average report a median of 9 symptoms over a week.[2] The Standardized Outcomes in Nephrology–HD study has identified outcomes that are important to physicians and patients.[3] Whereas cardiovascular disease and mortality outcomes are easily tracked and identified, symptoms are difficult to identify and usually require a prospective survey of patients or manual chart review, which are time consuming (with many surveys including more than 30 questions) and provide only 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) permits 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 patients undergoing HD. We then compared the performance of NLP and International Classification of Diseases (ICD) against manual chart review.

RESULTS

Patient characteristics

We identified 1080 patients receiving maintenance HD from the Bio*Me* biobank (Supplementary Figure S1A); 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. Patients had 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), with 342 progress notes (IQR, 102–782) and 16 discharge summaries (IQR, 2–54). The mean follow-up time was 8.7 ± 5.5 years (Table 1). From the Medical Information Mart for Intensive Care (MIMIC-III) database, we identified 519 patients undergoing chronic HD utilizing ICD-9 codes (Supplementary Figure S1B). The mean age of patients was 70 ± 39.6 years, 41% of patients were women, and 63% self-reported as European American. The prevalence of comorbidities was high, with diabetes at 54%, hypertension at 91%, coronary artery disease at 46%, and congestive heart failure at 47% (Table 1). The median progress note count was 10 (IQR 0–55), and the median discharge summary count was 1 (IQR, 1–2). Because a majority of patients only had 1 encounter, follow-up time could not be calculated.

Symptom identification using NLP versus administrative codes

In the Bio*Me* development cohort, NLP identified symptoms more frequently than did ICD codes (Figure 1a). 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 identified most frequently 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 frequencies 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 cramping, respectively. In the Bio*Me* validation cohort, the mean frequencies 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 identified 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 1b). Depression, anxiety, and pain were the most common symptoms identified 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 (κ statistic, 0.6–1). Frequency of symptoms was 4% to 54% as identified by NLP + ICD + manual review, 16% to 54% as identified by NLP + manual review, and 0 to 2% as identified by ICD + manual review (Figure 2a and b). Sensitivity for NLP ranged from 0.85 (95% confidence interval [CI], 0.65–96) for depression to 0.99 (95% CI, 0.93–1) for fatigue, whereas 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, whereas 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 3a; Supplementary Table S1A). ICD codes

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, yr	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, yr	8.7 ± 5.5	N/A[a]	6.9 ± 3.5

IQR, interquartile range; MIMIC-III, Medical Information Mart for Intensive Care; N/A, not available.
[a]Because a majority of patients only had 1 encounter, follow-up time was not calculated.
Data are shown as mean ± SD or count (%) except where specified.

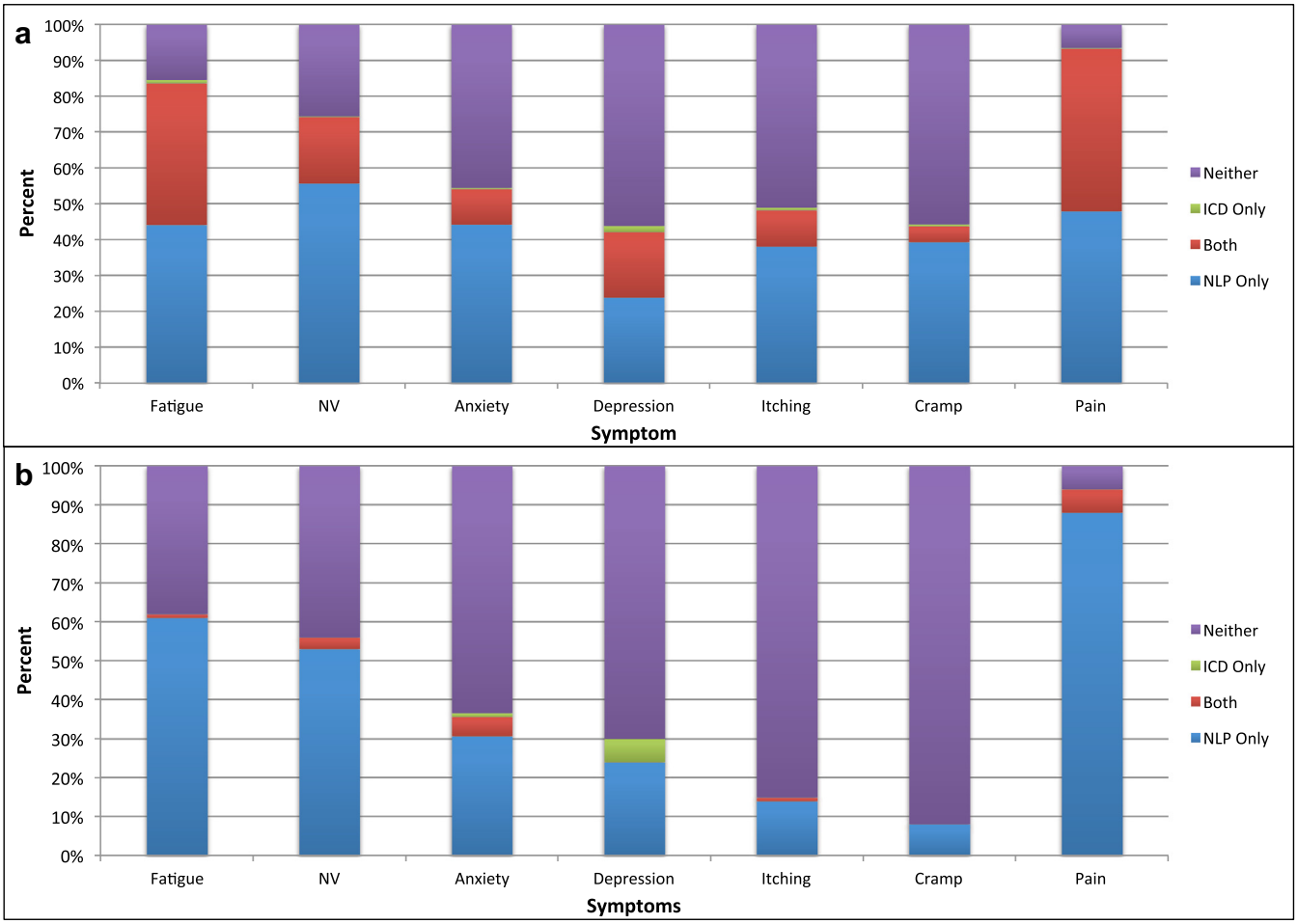

Figure 1 | Frequency of symptom identified by natural language processing (NLP) and International Classification of Diseases (ICD) from Bio*Me* (a) and Medical Information Mart for Intensive Care (MIMIC-III) (b). Blue bars indicate the percentage of patients in whom the symptom was found only by NLP, green bars indicate the percentage of patients in whom the symptom was found by only by ICD, red bars indicate the percentage of patients in whom the symptom was found by both NLP and ICD, and the purple bars indicate the percentage of patients in whom the symptom was found by neither NLP nor ICD. NV, nausea and vomiting.

were significantly more specific for N/V (NLP 0.57 [95% CI, 0.29–0.82] versus 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 but worsened specificity (Supplementary Figure S2A).

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 (Supplementary Table S1B).

Twenty-five patients were identified as having undergone Patient Health Questionnaire (PHQ)-9 depression screening, of which 24 patients were identified as having depression by PHQ-9 and/or clinical history. NLP correctly identified 22 patients (92%), whereas ICD-9/10 identified 20 patients (83%).

In MIMIC-III, sensitivity for NLP ranged from 0.5 for cramp to 0.98 (95% CI, 0.86–1) for fatigue, whereas sensitivity for ICD ranged from 0.04 (95% CI, 0–0.21) for fatigue to 0.5 (95% CI, not available) for cramp (Figure 2b; Supplementary Table S1C). Specificity for NLP ranged from 0.11 (95% CI, 0–0.48) for pain to 0.98 (95% CI, 0.89–1) for itching, whereas 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%), 61 (6%), 87 (8%), 99 (10%), 177 (17%), 158 (15%), 204 (20%), 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 per year, whereas patients with all 7 symptoms had a median of 24 (IQR, 15–43) encounters per year. A moderate correlation was found

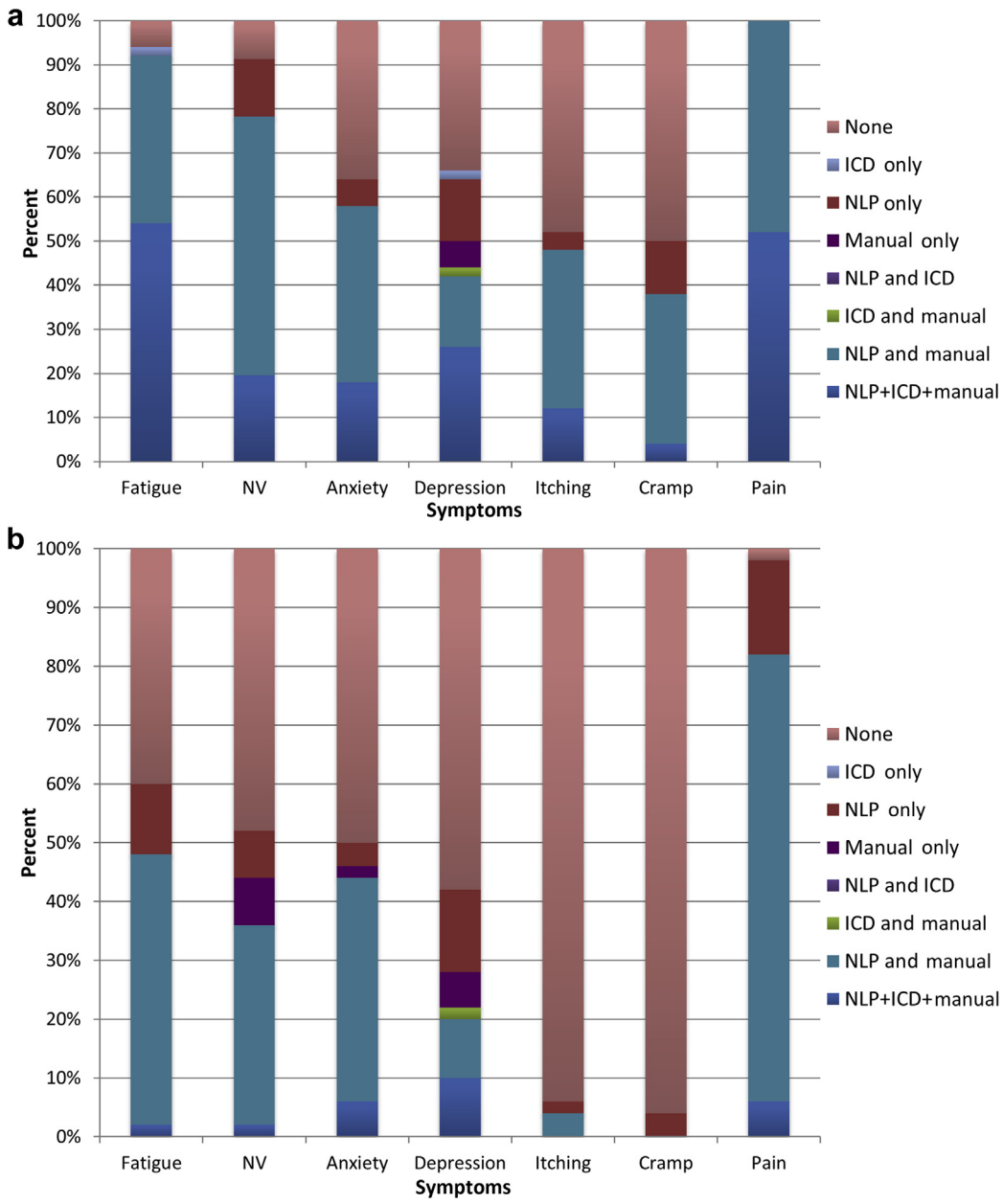

Figure 2| Frequency of symptoms from 50 patients from Bio*Me* (a) and Medical Information Mart for Intensive Care (MIMIC-III) (b) who had manual chart review. ICD, International Classification of Diseases; NLP, natural language processing; NV, nausea and vomiting.

between the number of encounters per year and the number of symptoms identified by NLP (correlation coefficient 0.36, $P < 0.001$). Among the 50 patients who had a manual chart review, symptom burden identified by NLP and manual chart review was similar; however, patients had fewer symptoms identified by ICD codes than by manual chart review (Figure 4a, b, and c). A moderate significant positive correlation was found 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%), 51 (10%), 136 (26%), 122 (24%), 98 (19%), 65 (13%), 21 (4%), and 5 (1%) patients with 0, 1, 2, 3, 4, 5, 6, and 7 symptoms, respectively.

Subgroup analysis in Bio*Me*

A total of 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 it did without date restrictions (fatigue 58% vs. 84%, N/V 39% 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 the 1-year subset vs. 93% with no restrictions). Even with date restrictions, NLP identified more symptoms and had better sensitivity than

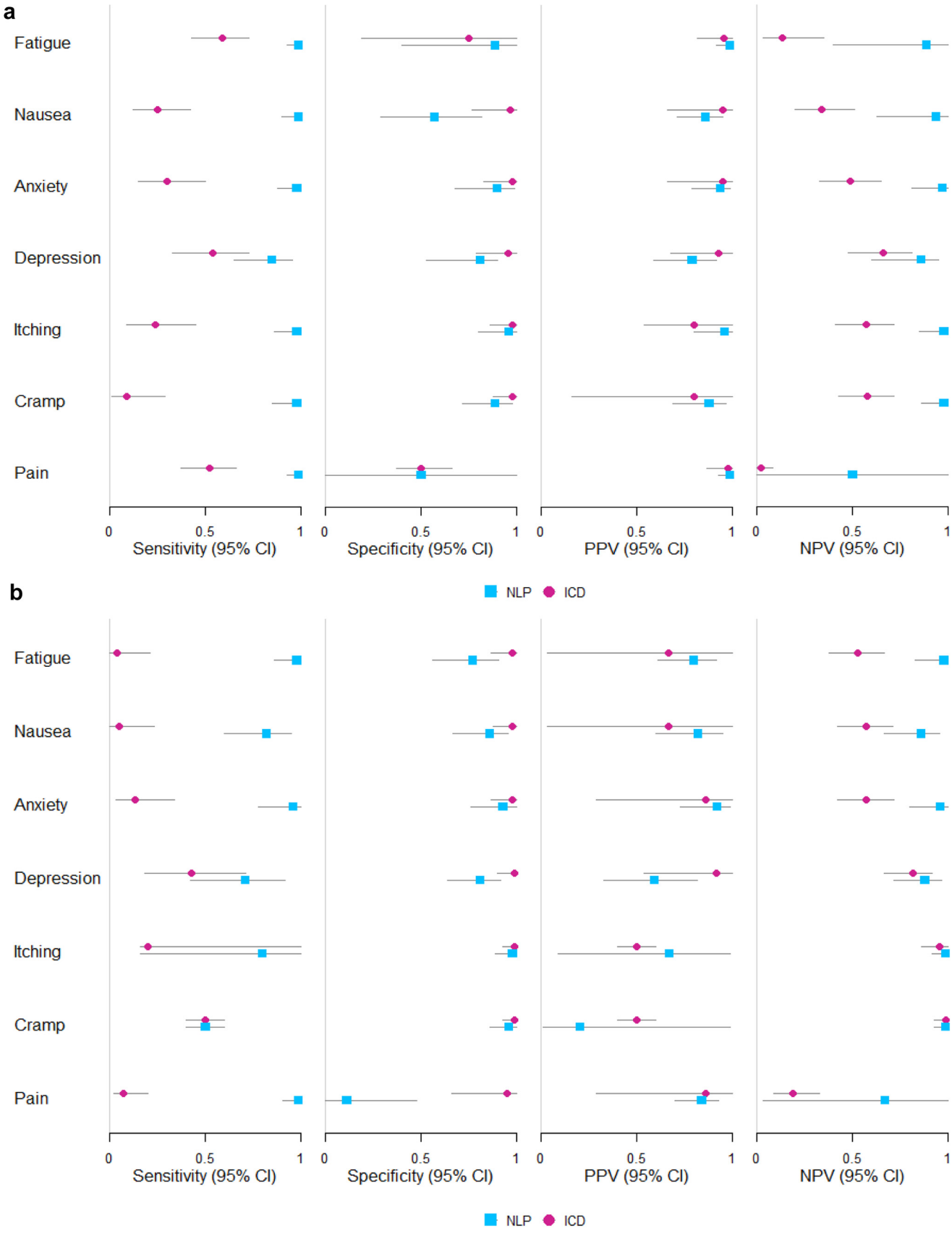

Figure 3 | Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of natural language processing versus International Classification of Diseases for the identification of symptoms for Bio*Me* (a) and Medical Information Mart for Intensive Care (MIMIC-III) (b) calculated using manual chart review of 50 patient charts. CI, confidence interval.

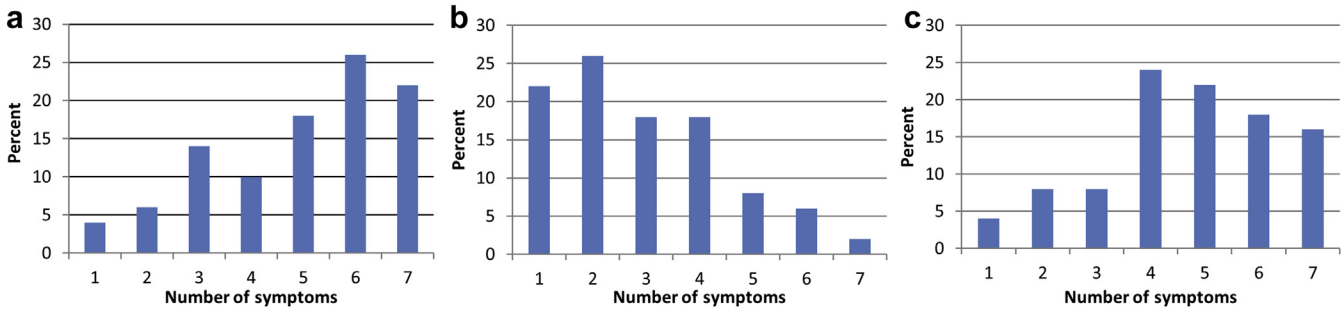

Figure 4 | Overall symptom burden demonstrating the number of symptoms identified from 50 Bio*Me* patients by natural language processing (a), International Classification of Diseases (b), and manual review (c).

did ICD codes (Supplementary Figure S3 and Supplementary Table S1D).

A total of 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 did ICD (Supplementary Figure S4).

Out of 100,118 notes, 11,066 (11%) were from an inpatient hospital stay. Symptoms were identified more frequently in outpatient notes than in inpatient notes except for N/V. Fatigue identification had the largest difference by inpatient and outpatient notes, with a difference of 19% (Supplementary Figure S5). Overall, NLP had better sensitivity for symptom identification in outpatient notes but better specificity in inpatient notes (Supplementary Figure S2B).

A total of 533 patients (52%) had 7476 episodes of pain with severity documented; 3137 (42%) were mild, 2232 (30%) were moderate, and 2107 (28%) were severe. NLP performed similarly across pain severity types (Supplementary Figure S2C).

DISCUSSION

Although symptoms are common in patients undergoing HD and are identified as important to patients and providers, efficient retrospective assessment of symptoms from the EHR is difficult. We show that NLP has better sensitivity than ICD codes at identifying 7 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 the number of encounters and number of symptoms identified by NLP.

The Standardized Outcomes in Nephrology–HD initiative identified several outcomes important to all stakeholders and has emphasized the importance of clinical research that includes these symptoms. Prior research that used patient-centered outcomes as end points 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. Although NLP can process notes in an efficient manner, few studies in nephrology 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.[2,8,15] Symptoms such as itching and cramps were less frequent, whereas other symptoms such as N/V were found more commonly. Differences in patients enrolled in studies and those seen in real-world practice likely contribute 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 compared with the general outpatient US HD population needs to be further elucidated.

Whereas surveys in prior studies have been conducted in patients who are stable at their outpatient HD centers, we included outpatient and acute inpatient notes. Subgroup analysis showed 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, whereas in the inpatient notes providers will focus on the admitting diagnosis. Because MIMIC-III consists of progress notes from critically ill patients, symptoms were identified at an even lower rate, likely because the patients were critically ill and unable to verbalize their symptoms and because 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 did ICD codes. Although 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 end-stage renal disease status. However, these symptoms remain important patient outcomes because they were deemed to be important to patients, physicians, and caregivers.[3]

We found that NLP outperformed ICD codes for symptom identification.[16–19] Because ICD codes are administrative and billing codes, clinicians may be less inclined to use them to document symptoms experienced by patients undergoing HD, especially if they do not count toward overall reimbursement. Sensitivity for ICD codes are generally moderate even in more common conditions such as myocardial infarction (72%) and hypertension (78%).[20] The addition of medications to ICD codes for identification of N/V, anxiety, depression, and pain increased sensitivity but decreased specificity, likely because of the use of medications for other indications (e.g., bupropion for depression and for smoking cessation).

ICD codes for symptoms had high specificity and high positive predictive value. Therefore, the NLP method may be favored for identification of a large cohort of patients with symptoms while accepting the risk of higher false-positive rates, whereas the ICD method may be favored for identification of patients highly likely to have symptoms while accepting the higher false-negative rate. Although NLP was more sensitive at identifying depression, ICD codes were more specific because of false-positives for depression used in other clinical contexts (e.g., depressions on electrocardiogram or temporal depressions). Although 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 because of 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, in which more data are available for both sicker patients and patients with longer length of follow-up.[21] Unfortunately, data regarding the authors of the notes are not available, and we cannot comment on the documentation of symptoms by provider type. Additionally, only symptoms that the provider is screening for are documented, and therefore NLP may miss symptoms that patients are not discussing with their providers, which may lead to an underestimation of symptom prevalence.[22] Neither the Bio*Me* nor MIMIC-III datasets are exclusive to patients undergoing HD on an outpatient basis, which makes comparison with prior published data difficult and reduces the generalizability of our results. However, the prevalence of symptoms in our study is similar to prior published survey data.[2,15] Data are extracted from EHRs of respective institutions, and this export process may affect the generalizability of results to other institutions. Although presence of symptoms varied throughout time, we chose to classify patients as ever present or never present because our goal was to evaluate the performance of NLP in identifying patients with the symptoms. In our subgroup analyses of 1 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, because 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 used NLP to identify important patient symptoms from the EHR of patients undergoing HD from the Mount Sinai health system and validated our results in MIMIC-III. NLP outperformed ICD codes for identification with regard to sensitivity, negative predictive value, 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 prior to using 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. We included patients undergoing HD, excluding those with a kidney transplant and those who never underwent dialysis. Because linkage information did not include dialysis type or dialysis access type, peritoneal patients were excluded using ICD codes (Supplementary Table S2A). 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 2 distinct cohorts: (i) the Bio*Me* validation cohort composed of patients undergoing chronic HD from Bio*Me* who were not included in the original development cohort and (ii) the MIMIC-III database.[23] The Bio*Me* validation cohort consisted of patients who enrolled in Bio*Me* between January 1, 2018, and March 15, 2019, and were identified using ICD-9/10 codes (Supplementary Table S2A). MIMIC-III is a critical care database of patients from a large, single center tertiary care hospital from 2001–2012.[23] We included all notes from the MIMIC-III database. Because MIMIC-III could not be linked with the United States Renal Data System, end-stage kidney disease 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 peritoneal patients by peritoneal dialysis codes (Supplementary Table S2A). Because MIMIC-III is a de-identified

publicly available database, evaluation of data from this source was considered exempt from institutional review board approval.

Patient comorbidities were identified using the Clinical Classification Software developed by the Healthcare Cost and Utilization Project.[24] The Clinical Classification Software aggregates ICD codes into clinically meaningful and mutually exclusive categories. The codes used for identification are included in Supplementary Table S2B.

Study design

This study is a retrospective cohort study of patients undergoing HD drawn from the EHRs of 2 medical systems. We utilized the CLiX NLP engine produced by Clinithink (London, UK) to parse notes. CLiX NLP is an NLP software that matches free text to Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT).[25] SNOMED CT is a comprehensive health care terminology resource that has an inherent hierarchy consisting of overarching concepts, that is, parent terms, which encompass more specific concepts, that is, children terms. Supplementary 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.[26,27] 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," and "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 the 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, itching, and cramps.[3] SNOMED CT for the associated outcomes were selected through extensive review by 2 physicians (Supplementary Table S3). These specific terms were selected because of 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 2 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, presence versus absence, and, if appropriate, location/laterality (Supplementary 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 the date of the note. Although CLiX NLP is a proprietary system, this study can be replicated with other NLP tools that utilize SNOMED CT. We chose this NLP method because of the research team members' familiarity with it and its availability to us. Other valid methods, including machine learning, were not used because of a lack of mature methodologies for this specific project.

We performed 2 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 article from the final NLP query with test statistics calculated from a separate manual chart review of 50 randomly selected charts as described later. Examples of false-positive results that were identified during this final chart review are presented in Supplementary Table S4. This final NLP algorithm was then validated in a cohort of distinct HD patients from Bio*Me* and MIMIC-III.

We compared performance of ICD–clinical modification (CM) codes with the results obtained from CLiX NLP.[28–30] ICD-9 and -10 codes were used in Bio*Me*, whereas only ICD-9 codes were available in MIMIC-III (Supplementary Table S2C). To determine if medication data improved ICD identification of symptoms, we identified medications used for pain, N/V, anxiety, and depression from RxNorm (Supplementary Table S5) and identified patients for whom these medications had ever been prescribed.[31] Medications that are commonly used for other indications (e.g., aspirin for second prevention of cardiac events) were removed from the list. Finally, both methods were compared with independent chart review by 2 physicians. We randomly selected 50 patient charts from Bio*Me* and MIMIC-III, using SAS (PROC SURVEYSELECT 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 control subjects (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 1 year of notes from patients who had at least 2 years of data. Manual chart review was done for 30 patients. Second, only patients with at least 1 encounter with notes available were included and grouped into tertiles based on the number of encounters (low, medium, and high). Unfortunately, the MIMIC-III database was solely an intensive care unit 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 PHQ-9 screening documentation.[32,33] We considered depression screening positive if patients scored $\geq$10 or there was evidence of a history of depression (i.e., cognitive behavior therapy, antidepressive medications, or prior suicide attempts).

Statistical analysis

We calculated sensitivity, specificity, positive predictive value, negative predictive value, and F1 scores of NLP and ICD-9/10 codes. F1 scores were calculated as a measure of accuracy that considers both the sensitivity and positive predictive value.[34] For cells on the 2$\times$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.[35,36] The 95% CIs were calculated using the PROC FREQ procedure in SAS using the binomial option.[37] We compared estimates of sensitivity and specificity using McNemar's test, with significance set using a 2-sided P value of <0.05. We compared negative predictive value and positive predictive value using the generalized score statistic method and the SAS macro created by Gondara *et al.*[38] 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's 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 (Copyright © 2019 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.) and R 3.6.0 (R Foundation for Statistical Computing, Vienna, Austria).

DISCLOSURE

GNN and SGC are co-founders of RenalytixAI, and GNN and SGC are members of the advisory board of RenalytixAI and own equity in the same. GNN has received operational funding from Goldfinch Bio. GNN has received consulting fees for BioVie Inc. SGC has received consulting fees from Goldfinch Bio, CHF Solutions, Quark Biopharma, Janssen Pharmaceuticals, and Takeda Pharmaceuticals. GNN and SGC are on the advisory board for pulseData and have received consulting fees and equity in return. TVV was part of launching Clinithink and retains a financial interest in the company. PK is an employee of the Renal Research Institute, a wholly owned subsidiary of Fresenius Medical Care (FMC); he holds stock in this company. All the other authors declared no competing interests.

ACKNOWLEDGMENTS

LC is supported in part by the National Institutes of Health (5T32DK007757–18). GNN is supported by a career development award from the National Institutes of Health (K23DK107908) and is also supported by R01DK108803, U01HG007278, U01HG009610, and U01DK116100. SGC is supported by the following grants from the National Institutes of Health: U01DK106962, R01DK106085, R01HL85757, R01DK112258, and U01OH011326.

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

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.

SUPPLEMENTARY MATERIAL

Supplementary File (Word)

Figure S1. Study flow chart for Bio*Me* (**A**) and MIMIC-III (**B**).
Figure S2. Test parameters for addition of medication on ICD performance (**A**), NLP performance for pain severity identification (**B**), and NLP performance on symptom identification (**C**) in inpatient versus outpatient notes.
Figure S3. Frequency of symptoms from Bio*Me* patients identified from 1 year of notes.
Figure S4. Frequency of symptoms identified by NLP and ICD across tertiles of encounters.
Figure S5. Proportion of symptoms identified in inpatient and outpatient notes.
Figure S6. SNOMED Hierarchy for concept of cramp.
Figure S7. Example of CLiX parsing the phrase "cramp in left leg."
Table S1. Sensitivity, specificity, PPV, NPV, and F1 score of NLP versus ICD for identification of symptoms for Bio*Me* development cohort (**A**); Bio*Me* validation cohort (**B**); MIMIC-III (**C**); and 1 year of notes from patients in Bio*Me* (**D**) calculated using manual chart review.
Table S2. ICD 9 and 10 codes for identification of hemodialysis and peritoneal dialysis patients (**A**); comorbidities (**B**); and symptoms (**C**).
Table S3. List of SNOMED concepts and children terms used for symptom identification.
Table S4. Example of false-positives and false-negatives found with NLP.
Table S5. List of medications for pain, nausea and/or vomiting, depression, and anxiety as per RXnorm.

REFERENCES

1. United States Renal Data System. 2018 USRDS annual data report: epidemiology of kidney disease in the United States. Bethesda, MD: National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 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.
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.
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.
5. Hays RD, Kallich JD, Mapes DL, et al. Development of the kidney disease quality of life (KDQOL) instrument. *Qual Life Res*. 1994;3:329–338.
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;33:1664–1671.
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 (Wash DC)*. 2016;4:1231.
8. Merkus MP, Jager KJ, Dekker FW, et al. 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;14:1163–1170.
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;22:872–880.
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.
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;17:588–594.
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;2014:907–916.
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.
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. Caplin B, Kumar S, Davenport A. Patients' perspective of haemodialysis-associated symptoms. *Nephrol Dial Transplant*. 2011;26:2656–2663.
16. 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;17:1688–1694.
17. 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.
18. Semins MJ, Trock BJ, Matlaga BR. Validity of administrative coding in identifying patients with upper urinary tract calculi. *J Urol*. 2010;184:190–192.
19. McCormick N, Lacaille D, Bhole V, et al. Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis. *PLoS One*. 2014;9:e104519.

20. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care*. 2005;43:1130–1139.
21. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc*. 2013;2013:1472–1477.
22. Weisbord SD, Fried LF, Mor MK, et al. Renal provider recognition of symptoms in patients on maintenance hemodialysis. *Clin J Am Soc Nephrol*. 2007;2:960–967.
23. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
24. Agency for Healthcare Research and Quality. Healthcare Cost and Utilization Project (HCUP). www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed November 11, 2019.
25. Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. *Proc AMIA Annu Fall Symp*. 1997:640–644.
26. 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.
27. 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.
28. 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.
29. 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.
30. Kisely S, Lin E, Gilbert C, et al. Use of administrative data for the surveillance of mood and anxiety disorders. *Aust N Zeal J Psychiatry*. 2009;43:1118–1125.
31. Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, and Department of Health & Human Services. RxClass. Exploring Classes for RxNorm Drugs. Available at: https://mor.nlm.nih.gov/RxClass. Accessed October 25, 2019.
32. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann*. 2002;32:509–515.
33. 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.
34. Sasaki Y. The truth of the F-measure. https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf. Published 2007. Accessed November 11, 2019.
35. Lawson R. Small sample confidence intervals for the odds ratio. *Commun Stat Simul Comput*. 2004;33:1095–1113.
36. Dureh N, Choonpradub C, Tongkumchum P. An alternative method for logistic regression on contingency tables with zero cell counts. https://rdo.psu.ac.th/sjstweb/journal/38-2/38-2-8.pdf. Published 2016. Accessed November 11, 2019.
37. SAS. 24170–Estimating sensitivity, specificity, positive and negative predictive values, and other statistics. http://support.sas.com/kb/24/170.html. Accessed November 11, 2019.
38. Gondara L. *A SAS® macro to compare predictive values of diagnostic tests*. https://www.researchgate.net/publication/281440778_A_SAS_R_macro_to_compare_predictive_values_of_diagnostic_tests. Published April 2015. Accessed November 11, 2019.