eGEMs
Generating Evidence & Methods to improve patient outcomes

Enhanced Quality Measurement Event Detection: An Application to Physician Reporting

Suzanne R. Tamang, PhD; Tina Hernandez-Boussard, MS, PhD, MPH; Elsie Gyang Ross, MD, MsC; Gregory Gaskin, MD; Manali I. Patel, MPH; Nigam H. Shah, MBBS, PhD[i]

ABSTRACT

The wide-scale adoption of electronic health records (EHR)s has increased the availability of routinely collected clinical data in electronic form that can be used to improve the reporting of quality of care. However, the bulk of information in the EHR is in unstructured form (e.g., free-text clinical notes) and not amenable to automated reporting. Traditional methods are based on structured diagnostic and billing data that provide efficient, but inaccurate or incomplete summaries of actual or relevant care processes and patient outcomes. To assess the feasibility and benefit of implementing enhanced EHR-based physician quality measurement and reporting, which includes the analysis of unstructured free-text clinical notes, we conducted a retrospective study to compare traditional and enhanced approaches for reporting ten physician quality measures from multiple National Quality Strategy domains. We found that our enhanced approach enabled the calculation of five Physician Quality and Performance System measures not measureable in billing or diagnostic codes and resulted in over a five-fold increase in event at an average precision of 88 percent (95 percent CI: 83-93 percent). Our work suggests that enhanced EHR-based quality measurement can increase event detection for establishing value-based payment arrangements and can expedite quality reporting for physician practices, which are increasingly burdened by the process of manual chart review for quality reporting.

[i]Stanford University

Introduction

About 30 percent, or 117 billion dollars, of Centers for Medicare and Medicaid Services (CMS) payments are now linked to quality of care delivery, with the goal of linking 90 percent payments to quality of care by 2020.[1] As a result, a new framework for rewarding health care providers for provision of high value care is needed in the United States. However, there have been substantial obstacles in establishing efficient and meaningful quality reporting systems and pay for performance programs, leading to increasing concerns from professional groups and health policy experts.[2-22]

One salient challenge for physician quality reporting systems is the significant gap between what is desirable to measure for establishing value-based payment arrangements, and what is feasible for practices to report about the quality of care they deliver.[2,3,5,6,8,11,14-16,21,23-27] Although the wide-scale adoption of electronic health records (EHRs) has increased the availability of routinely collected structured (e.g., ICD diagnoses and CPT codes), and unstructured (e.g., free-text clinical notes) data that can be used to improve the reporting of quality of care, the bulk of information in the EHR is in unstructured and not amenable to automated reporting.

Traditional quality measurement and reporting methods consider only structured EHR data. The high degree of organization facilitates automated reporting, but structured data provide a limited representation of each patient's treatment across settings, their health outcomes, or why a clinical decision was made (e.g., a guideline was not followed because patients could not tolerate a high dose of medication).[12,13,23,28,29] Unstructured EHR data, such as the free-text notes that a clinical team documents as part of a patient's care process, is the opposite and represents a rich and complex pool of clinical information that does not conform to a pre-specified format. Although traditional methods for performance and quality reporting miss a substantial amount of information that is "locked" in clinical text, to warrant their wide scale adoption for population health management and measurement, clinical text analysis tools must be able to efficiently analyze an institution's entire clinical text collection, easily adapt to new information extraction tasks, and demonstrate reliable performance across institutions and different EHR products.

Information extraction is the task of automatically extracting structured information from unstructured data sources. We developed and evaluated information extraction methods to assess the feasibility of and potential benefit of improving EHR-based measurement for the Physician Quality and Performance Reporting System (PQRS). The PQRS is a physician quality reporting program implemented by the Centers for Medicare and Medicaid Services (CMS), under the Tax Relief and Health Care Act of 2006.[30] In combination with the cost of care an eligible health professional delivers, the PQRS is used by Medicare as the basis for differential payments based on healthcare "value".[7,18,30,31]

We discuss our research findings in the context of value-based payment for physicians. To the best of our knowledge, our work is the first systematic comparison of enhanced EHR-based methods with traditional methods in the setting of a physician performance and quality measurement. Prior studies applying clinical text analysis for enhanced measurement largely reflect research on improving event detection for the patient safety domain, such as the adverse events measured by Agency for Healthcare Research and Quality (AHRQ)'s Patient Safety Indicators (PSIs) and other surgery-related complications.[32-36] We sought to assess the feasibility and benefit of enhanced measurement, spanning multiple National Quality Strategy domains

and healthcare settings. We also depart from prior work in the use of a hybrid framework for clinical text analysis, CLEVER (from CLinical EVEnt Recognizer).[37] CLEVER is an open source tool that incorporates statistical term expansion components (i.e., word embedding) and semantic components (i.e., context analyzing rules) to expedite the development of rule-based extractors.[38,39]

Methods

Participants and Setting

Stanford Health Care (SHC) is an academic medical center located in Northern California. SHC provides inpatient and outpatient care for patients with high acuity disease with a recent focus on primary care. During the time of our analysis, SHC used the Epic (Epic Systems, Verona Wisconsin) EHR. Among the 646,973 patients that received care at SHC from 2008 through 2013, 178,794 senior patients, 4,213 dementia patients, 2,335 cataract surgery patients, and 7,414 ischemic stroke patients satisfied one or more of the ten PCPI denominator definitions. The SHC data for our study was provided by the STRIDE (Stanford Translational Research Integrated Database Environment), which contains deidentified data for over 2 million patients.[40] We analyzed over 21 million notes in the Epic EHR, including Letters, Phone Encounter Logs, Goals of Care and more standard note types such as Progress Notes, Nursing Sign Out Notes, ED Notes and other types.

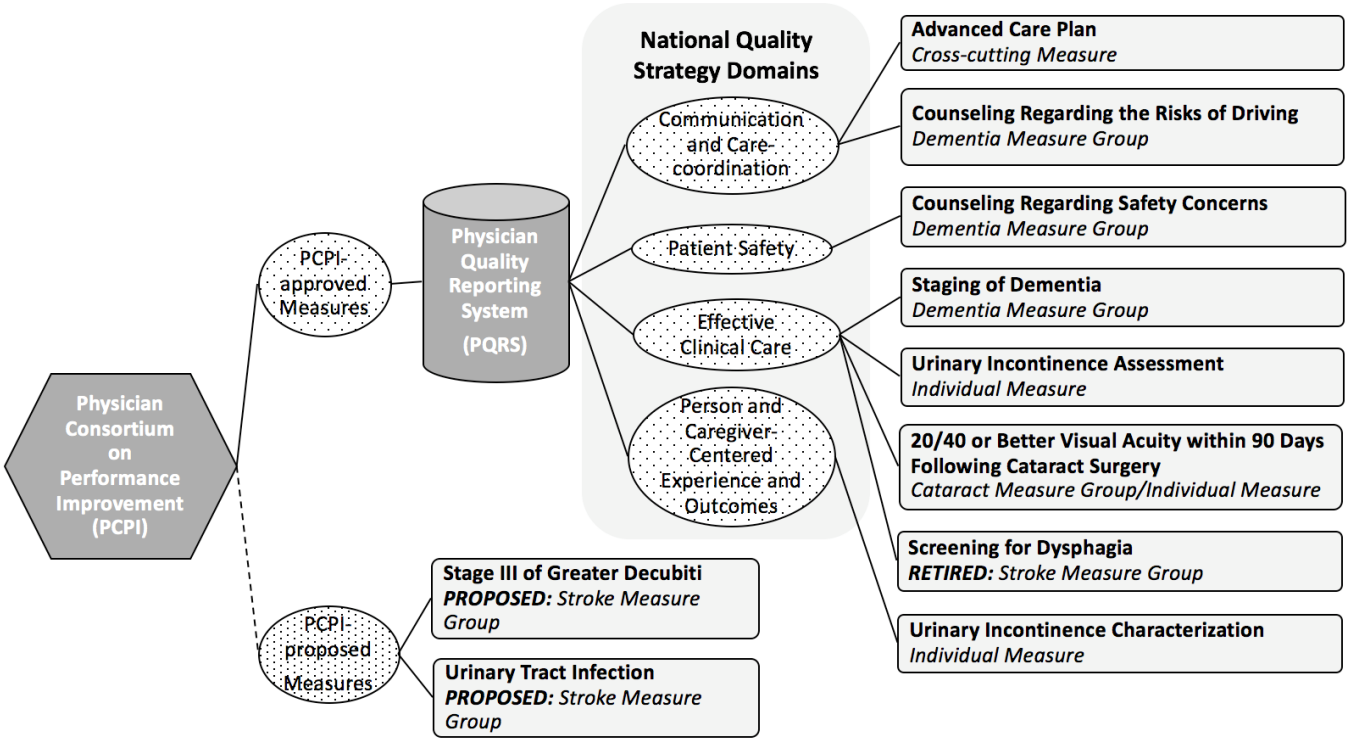
Physician Measures

Our study included ten physician performance and quality of care measures developed and approved by the Physician Consortium for Performance Improvement (PCPI), convened by the American Medical Association (AMA). Along with their name, PCPI group, and quality domain, each of our ten study measures are categorized by PCPI approval and PQRS adoption status in Figure 1.

The measure technical specifications determined by PCPI standardize the collection of measures and are distributed, without modification, for claims and registry based reporting by CMS, private companies, and professional groups.[41, 42] Each measure's technical specification includes the Measure Description, Measure Components, Measure Importance and Measure Designation. The Measure Description provides a short description of the measure; for example, the Measure Description for PQRS Measure #48, Urinary Incontinence Assessment, is "Percentage of female patients aged 65 years and older who were assessed for the presence or absence of urinary incontinence within 12 months". The Measure Components refer to the *Denominator Statement* and the *Numerator Statement*. For example, the Measure Components for Measures #48, Urinary Incontinence Assessment, indicates the age, gender and CPT or HCPCS codes for identifying patients to include in the measure's denominator; also, the set of CPT codes for calculating the numerator. The Measure Importance describes the relevance of each measure, and the Measure Designation categorizes each measure by type and National Quality Strategy domain. We provide the Measure Description and the Numerator and Denominator Statements from the PCPI Measure Component section in our Appendix, Table A1.

Using PCPI technical specifications, we used only structured data to estimate patient eligible for the denominator of each study measure. Keeping the denominator's value fixed, we compared traditional structured and enhanced methods for estimating the numerator. We prioritized the choice of measures by identifying measures that we hypothesized would be under-coded in structured EHR fields. For example, a measurement approach that used structured data for the calculation of Measure #47, Advance Care Plan, could identify events relevant to the numerator

Figure 1. Physician Quality Measures by TITLE, National Quality Strategy Domain, Measure Type and PCPI Approval Status



by detecting a specific CPT code, but most providers are unaware of them and many relevant discussions between patients and physicians, and advance care plans, are still documented in only free-text.

After recognizing that some measure numerators cannot be represented by the PQRS core coding systems (e.g., ICD, CPT, HCPCS), we also included measures that required disease recognition at a level of specificity that current coding systems do not support. For example, PQRS Measure #280, Staging of Dementia, is defined by the percent of Dementia patients that were staged as mild, moderate, or severe. The ICD allows for more granular dementia diagnoses such as presenile, senile, vascular, and drug induced, but cannot capture information on whether a patient's dementia is mild, moderate, or severe. Similarly, the PQRS core coding systems cannot be used to report Measure #191, "the percentage of patients aged 18 years and older with a diagnosis of uncomplicated cataract who had cataract surgery and no significant ocular conditions impacting the visual outcome of surgery and had best-corrected visual acuity of 20/40 or better (distance or near) achieved within 90 days following the cataract surgery patient reported 20/40 or greater vision within 90 days of cataract surgery". Although International Classification of Diseases Version 10 Clinical Modification (ICD10-CM) allows for the indication of laterality, it has no code that can be used to indicate a patient with 20/40 or better vision.

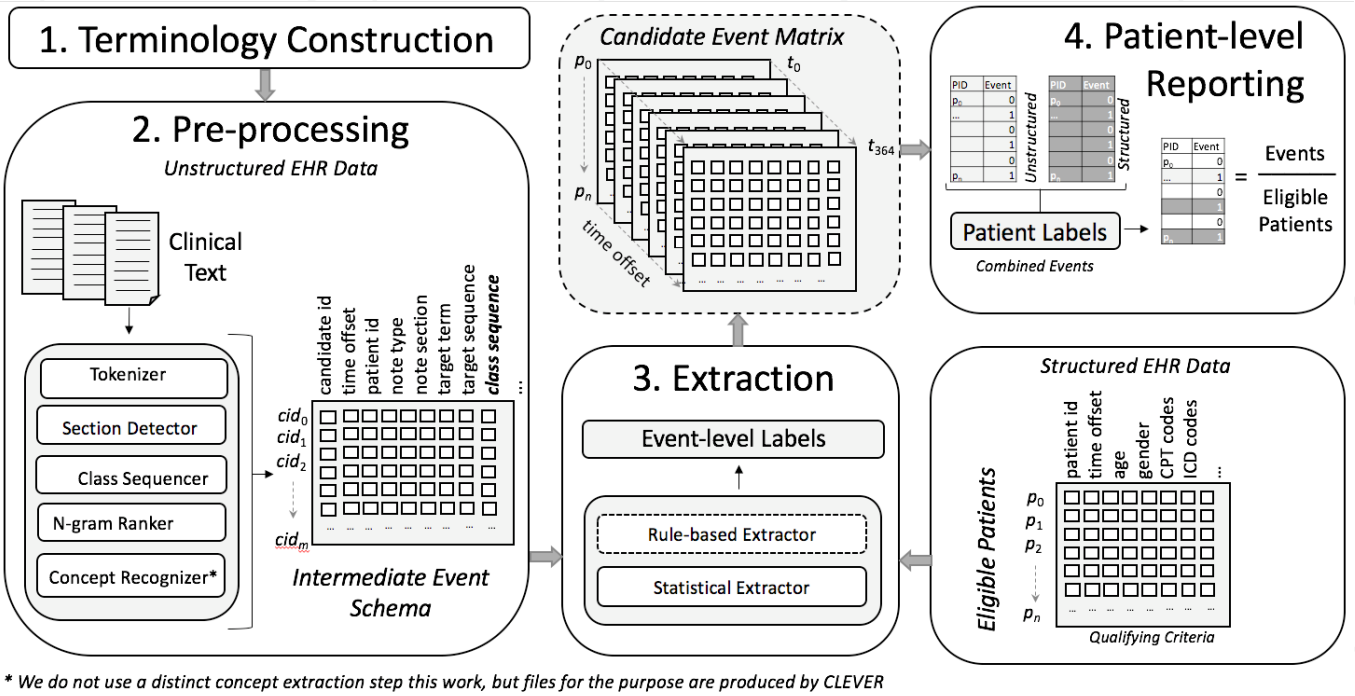
Clinical Information Extraction

Our approach to the detection of events "locked" in clinical text was to employ an efficient and flexible framework for building custom extractors, called *CLEVER*. Our clinical information extraction system consisted of four steps. Figure 2 shows an overview of our system's pipeline, based on a set of n patients, where p_i is the ith patient and $1 \leq i \leq n$,m candidate events, where cid_j is the jth candidate event and $1 \leq j \leq m$ and a measurement observation window, t, of one year, $t = 365$.

Step One of our information extraction pipeline was *Terminology Construction*. This step involved identifying the target concepts associated with the documentation of each measure in clinical text. Then, after using the UMLS and the SPECIALIST Lexicon to identify a set of high-quality biomedical "seed" terms for our target concepts, we used statistical term expansion techniques to identify new clinical terms that shared the same contexts.[43-45]

Step Two, *Pre-Processing*, used our terminology to tag the target terms for our measure-specific target concepts and a general set of clinical concept modifiers included in CLEVER. For each target term tagged in clinical text, CLEVER extracted a range of high and low-level features such as the term's target class, the surrounding context, or "snippet", entailing the target term, the note section, creation time and type (e.g., outpatient progress note), and the patient's ID. After the features were populated into CLEVER's *intermediate event schema*, we executed CLEVER's rule-based labeling algorithm and added a column to the intermediate event schema, indicating a positive or non-positive label for each candidate event.

Figure 2. Overview of the Clinical Information Extraction Pipeline for Enhanced EHR-based Reporting



** We do not use a distinct concept extraction step this work, but files for the purpose are produced by CLEVER*

In Step Three, *Extraction*, we used structured EHR data to estimate each measure's Denominator and Numerator Statements. By matching on patient ID and year, we merged data from our structured EHR data sources with CLEVER's intermediate event schema to create a *candidate event matrix*. We used note creation time to approximate the time of positive events extracted from clinical text, or in the case of events documented in structured data, visit time. Then, based on the approximated event time, all patients in the measures denominator were indexed from t_1, the date of their initial qualifying visit, through t_{max}, where "max" was the length of the measurement period indicated by each measure's PCPI technical specification. In the final step, *Patient-Level Reporting*, we queried the candidate event matrix to calculate the value of each measure's numerator and denominator, reporting the final measurement rate. CLEVER and the terminologies used for our experiments are publically available and distributed as open source software under the MIT license. Additional details on our information extraction pipeline illustrated in Figure 2 are provided in our Appendix.

Evaluation

The main study outcomes we report are the measurement rates for traditional and enhanced EHR-based quality measurement, and the precision, or PPV, of enhanced event detection from clinical text.

We evaluated the precision of our enhanced quality measurement method, using patients that qualified for the Numerator Component, based on only unstructured EHR data. For each study measure, we randomly selected 100 patients that were potentially missed by traditional measurement methods, for review by clinical experts. If the reviewer felt enough evidence was present in a patient's record to support the inclusion of the patient in the numerator, the clinical experts were instructed to indicate ***true***. For instances where there was a contraindication or insufficient evidence, they were instructed to indicate ***false***. To facilitate their case review, we included information from CLEVERs intermediate event schema – e.g., the snippets for target terms, the note type the target term appeared, the note's ID (NID), the time of the note, and the patient's ID. An example of four events that were selected for evaluation and ground truth (GT) labels that were assigned by clinical experts appear in Table 1.

Results

For our ten PQRS measures, we extracted seven measures from unstructured EHR data -- i.e., clinical text, -- with 80 percent or higher precision. For the seven measures enhanced measurement extracted with good performance, Table 2 shows the number of patients with clinical events that qualified for inclusion in each measure's numerator in the "Patient Events" column. The number of patients in the denominator appears under "Eligible Patients." On the left side of the "Patient Events" column, under "Structured Events," are the total number of patients in the numerator identified with traditional measurement. Under "Unstructured Events" is the number of additional patients identified with enhanced measurement methods. We also show the precision of enhanced measurement for the Numerator Component, and in the last two columns of Table 2, we compare traditional and enhanced measurement rates, based on six years of annual reporting (2008 though 2013).

For the seven PQRS measures in Table 2, a total of 384,503 patients contributed to measure denominators, from 2008 to 2013. Among all patients eligible for PQRS reporting, traditional claims-based reporting methods identified 1,580 patients for inclusion in measure numerators. We found that our enhanced quality measurement

Table 1. Evaluation Examples

GT	SNIPPET	NTYPE	PID	NID	TIME
MEASURE #47 ADVANCE CARE PLAN					
T	Transition to comfort care HPI XX yo F from SNF Sent to ED for increased WOB. Per daughter and granddaughter (DPOA) pt would not want anything done (including fluids antibiotics meds) and would like to be allowed to pass peacefully.	History and Physical	4325	346	XXX
T	Discussed goals of care with pt and her husband. Pt brought in her advance directive which names her husband [name ommited] as her surrogate decision maker. However, she has not documented her wishes with regards to life prolonging measures.	History and Physical	6341	645	XXX
MEASURE #191: 20/40 OR BETTER VISUAL ACUITY WITHIN 90 DAYS FOLLOWING CATARACT SURGERY					
T	The patient is also alert and oriented to person place and time. Distance Visual Acuity Right Eye Without correction With correction 20/40 -1 With Pin-Hole Autorefraction x Left Eye Without correction With correction 20/25 -1 With Pin-Hole Autorefraction x 3	Letter	3455	166	XXX
F	Distance Visual Acuity Right Eye Without correction CF at 3' With correction With Pin-Hole NI Autorefraction x Left Eye Without correction 20/80 With correction With Pin-Hole Autorefraction x 3 Intraocular Pressure	Progress Note, Outpatient	4425	169	XXX

Notes: For PQRS Measure #47, Advance Care Plan and Measure #191, 20/40 or Better Visual Acquity within 90 Days following Cateract Surgery shown by expert asigned ground truth label (GT), snippet, candidate event ID (CID), patient ID (PID), note ID (NID) and note type (NTYPE).

approach identified an additional 13,914 patients missed by traditional methods with an average precision of 88 percent (95 percent CI: 83-93 percent). These additional events improved the assessment of PQRS measures spanning four different National Quality Strategy domains – i.e., *Care-Coordination, Patient Safety, Effective Clinical Care, and Person and Caregiver Experience and Outcomes*.

For the two study measures that traditional method detected from ICD codes —Measure #48 Urinary Incontinence Assessment and Measure #49 Urinary Incontinence Characterization— enhanced quality measurement methods resulted in approximately an eight to and four-fold increase in the number of patients that satisfied the Numerator Statement of each measure, respectively. In addition, based on a traditional quality measurement approach, five PQRS

Table 2. Total Patient Events, Text-Based Extraction Precision and MULTI-YEAR Measurement RATES FOR PQRS MEASURES Using Traditional and Enhanced Quality Reporting

PQRS MEASURE	PATIENT EVENTS		PRECISION (PPV)	ELIGIBLE PATIENTS: *DENOMINATOR*	TRADITIONAL RATE (%): *TRADITIONAL NUMERATOR* (STRUCTURED EVENTS)/ *DENOMINATOR*	ENHANCED RATE (%): *ENHANCED NUMERATOR* (STRUCTURED EVENTS + UNSTRUCTURED EVENTS)/ *DENOMINATOR*
	STRUCTURED EVENTS	UNSTRUCTURED EVENTS	TEXT-BASED EXTRACTION			
Measure #47 (NQF 0326): Advanced Care Plan	0	2412	0.92	181734	0.00	1.33
Measure #48: Urinary Incontinence Assessment of Presence or Absence	1002	7999	0.91	178794	0.56	5.03
Measure #49: Urinary Incontinence Characterization	578	2320	0.98	9001	6.42	32.20
Measure #191 (NQF 0565): 20/40 or Better Visual Acuity within 90 Days Following Cataract Surgery	0	423	0.76	2335	0.00	18.12
Measure #280: Staging of Dementia	0	574	0.89	4213	0.00	13.62
Measure #286: Counseling Regarding Safety Concern	0	144	0.82	4213	0.00	3.42
Measure #287: Counseling Regarding Risks of Driving	0	42	0.92	4213	0.00	1.00

measures reported zero patients in their numerator. As shown in Table 2, our enhanced measurement approach enabled PQRS measurement and reporting for the following PQRS measures: Staging of Dementia (574), Counseling Regarding Safety Concern (144), Counseling Regarding Risks of Driving (42), Advanced Care Plan (2,412) and 20/40 of Better Visual Acuity within 90 days Following Cataract Surgery (423).

To examine annual reporting trends, based on our enhanced measurement method, Table 3 show shows annual reporting rates. Similar to Table 2, the numerator and denominator of each measure appear in the "Patient Events" and "Eligible Patients" columns, respectively. To quantify the change in physician performance between consecutive reporting year, and beginning with 2009, the column "Annual Improvement" reflects the relative difference in the annual measurement rate between consecutive years. For example, Measure #280, Staging of Dementia, showed an almost 1.61 percent improvement in performance between 2010 and 2011.This measure continued to improve by another 1.67 percent between 2011 and 2012. None of our physician performance and quality measures consistently increased over the six-year period and overall; we observed incremental improvements in the PQRS study measures over the six-year period and no dramatic decreases in reporting rates between consecutive years.

Discussion

Quality measurement and reporting is evolving to accommodate a more comprehensive definition of quality in healthcare delivery. However, the gap between what is desirable to measure and what is possible for providers to report on the quality of care they deliver is significant. Given the critical role quality reporting systems have in the success of value-based reform, our work suggests that fundamental changes in EHR-based data analysis and automated reporting software are required to support value-based care.

Other studies of enhanced EHRs, although using somewhat different methodologies, have shown that unstructured EHR data provides a wealth of rich information that can be used for quality reporting. A key innovation of our work is the application of clinical text analysis to physician performance and quality reporting. We found that enhanced quality measurement and reporting systems have the potential to improve physician quality reporting systems in several important ways. First, in conjunction with information on the cost of care they deliver, the ability to increase event detection across multiple National Quality Strategy domains provides more accurate and comprehensive information for establishing value-based spending arrangements. Whereas other studies focus on the detection of patient safety, we demonstrate the ability of enhanced measurement to extract 13,914 *additional* events that were missed by traditional methods, spanning four different National Quality Strategy domains: Care-coordination, Patient Safety, Effective Clinical Care and Person and Caregiver Experience and Outcomes.

Second, we found that enhanced quality reporting methods can enable physician practices to calculate quality measures which are associated with overall patient care, including coordination of care and patient satisfaction. Although the role of the individual in healthcare is changing from that of a passive patient to active consumer with increased financial responsibility for their healthcare costs, current quality reporting systems overwhelmingly focus on process measures that capture medical aspects of a patient's care (e.g., cancer screening according to guidelines, or the administration of prophylaxis before surgery). To provide patients with meaningful information to compare costs and

Table 3. PQRS Measurement Rates by Year as Measured by Enhanced Quality Measurement Reporting

PQRS MEASURE	MEASUREMENT YEAR	STRUCTURED AND UNSTRUCTURED EVENTS: *ENHANCED NUMERATOR*	ELIGIBLE PATIENTS: *DENOMINATOR*	ENHANCED MEASUREMENT RATE (%)	ANNUAL IMPROVEMENT (%)
Measure #40: Advanced Care Plan NQF(0326)	2008	148	22146	0.67	—
	2009	250	26297	0.95	0.28
	2010	380	31836	1.19	0.24
	2011	519	34986	1.48	0.29
	2012	695	40684	1.71	0.23
	2013	*420*	*25785*	*1.63*	*-0.08*
Measure #48: Urinary Incontinence Assessment	2008	980	23508	4.17	—
	2009	1331	27516	4.84	0.67
	2010	1578	30305	5.21	0.37
	2011	1777	33443	5.31	0.10
	2012	*1972*	*39018*	*5.05*	*-0.26*
	2013	1363	25004	5.45	0.40
Measure #50: Urinary Incontinence Characterization	2008	316	23508	1.34	—
	2009	452	27516	1.64	0.30
	2010	524	30305	1.73	0.09
	2011	578	33443	1.73	0.00
	2012	*601*	*39018*	*1.54*	*-0.19*
	2013	427	25004	1.71	0.17
Measure #191: Cataracts - 20/40 or Better Visual Acuity within 90 Days Following Cataract Surgery (NQF 0565)	2008	25	384	6.51	—
	2009	38	379	10.03	3.52
	2010	96	482	19.92	9.89
	2011	88	393	22.39	2.47
	2012	*100*	*460*	*21.74*	*-0.65*
	2013	76	237	32.07	10.33
Measure #280: Dementia Measure Group - Staging of Dementia	2008	59	473	12.47	—
	2009	74	576	12.85	0.38
	2010	*95*	*792*	*11.99*	*-0.86*
	2011	117	860	13.60	1.61
	2012	155	1015	15.27	1.67
	2013	*74*	*497*	*14.89*	*-0.38*

Notes: Measure, reporting year, the number of patient events (measure's numerator), eligible patients (measure's denominator), enhanced measurement rate (unstructured and structured data sources), and the Annual rate of change from the prior year Are indicated in each column. Years that show a decrease in the institutional annual perfromance of a quality measure appear in italics.

Table 3. PQRS Measurement Rates by Year as Measured by Enhanced Quality Measurement Reporting (Cont'd)

PQRS MEASURE	MEASUREMENT YEAR	STRUCTURED AND UNSTRUCTURED EVENTS: *ENHANCED NUMERATOR*	ELIGIBLE PATIENTS: *DENOMINATOR*	ENHANCED MEASUREMENT RATE (%)	ANNUAL IMPROVEMENT (%)
Measure #286: Dementia Measure Group - Counseling Regarding Safety Concerns	2008	4	23508	0.85	—
	2009	13	27516	2.26	1.41
	2010	29	30305	3.66	1.40
	2011	*26*	*33443*	*3.02*	*-0.64*
	2012	47	39018	4.63	1.61
	2013	25	25004	5.03	0.40
Measure #287: Dementia Measure Group - Counseling Regarding Risks of Driving	2008	5	23508	1.06	—
	2009	*5*	*27516*	*0.87*	*-0.19*
	2010	8	30305	1.01	0.14
	2011	10	33443	1.16	0.15
	2012	8	39018	0.79	-0.37
	2013	6	25004	1.21	0.42

Notes: Measure, reporting year, the number of patient events (measure's numerator), eligible patients (measure's denominator), enhanced measurement rate (unstructured and structured data sources), and the Annual rate of change from the prior year Are indicated in each column. Years that show a decrease in the institutional annual perfromance of a quality measure appear in italics.

quality among providers, payment arrangements linked to quality of care must incorporate value from the perspective of multiple stakeholders, including patients. For example, our enhanced quality measurement approach resulted in over a four-fold increase in one PQRS measure from the Patient and Caregiver Outcome and Experience domain of the National Quality Strategy. In addition, enhanced quality measurement enabled the quantification of five study measures that could not be reported using traditional methods, including two PQRS measures from the Communication and Care-coordination domain.

A third opportunity enabled through advancements in enhanced EHR-based systems is reporting efficiency. As of December 2015, over 450,000 eligible professionals chose a negative payment adjustment instead of participating in the PQRS program.[2,7] Low participation has been attributed to a complex and labor-intensive measurement process, which has been estimated to cost over $40,000 per practice and over 14 billion dollars across all practices in 2015.[2,7] Without a "computable" quality measurement and reporting infrastructure for the whole EHR, including clinical text, physician quality reporting will continue to demand a costly and labor intensive manual chart abstraction. For example, all three measures from our Dementia Measure Group were detected *only* by our enhanced measurement approach. Even with a state-of-the-art EHR, a physician practice reporting the Dementia Measure Group must engage in a burdensome administrative process to satisfactorily participate in the 2016 PQRS program.

It is important to note that our study has several limitations. Not all measures were equally amenable to our enhanced measurement method. Specifically, we do not report measurement rates for the three PCPI measures from the Stroke and Rehabilitation group, shown in Figure 1. Based on both traditional and enhanced measurement methods, the dysphagia measure, which is a retired PQRS measures, resulted in zero patients in the numerator and was omitted from our results. For the two Potentially Preventable Harmful Event measures from this group, our enhanced measurement method detected positive conditions of UTI or Stage 3 or greater decubitus, but could not rule out the absence of the conditions on hospital admission and the precision was 10 percent and 14 percent, respectively. Since such low performance is unlikely to reflect meaningful results, and the measure has not been adopted by the PQRS, we also omitted these measures from our results.

The lack of a gold standard corpus to evaluate our clinical text analysis methods is also a limitation of our study. We demonstrate the benefit of enhanced quality reporting by comparing traditional and enhanced EHR-based measurement approaches, and showing substantial increases in the total number of additional events detected from unstructured sources, with good precision. However, the recall (or sensitivity) of our system is unknown. Similar to open domain information extraction tasks, the size and complexity of our corpus makes it infeasible to manually annotate a sample of clinical notes that is large enough to provide a meaningful and unbiased estimate of system recall. A community resource of this type would be invaluable for developing and evaluating enhanced EHR-based measurement and quality reporting systems. Finally, a multi-site evaluation is needed to establish the generalizability of our methods.

Conclusions

For ten physician quality measures, we compared traditional reporting methods that considered only structured EHR data (e.g., diagnosis and procedural codes) with an enhanced EHR-based approach that included clinical text analysis. Based on our analysis of six years of EHR data from patients who visited Stanford Health Care, we found a total of 13,914 additional patients encounters relevant to the Numerator Component of measures adopted by Medicare's Physician Quality Reporting System and identified relevant clinical events with good precision (88 percent; 95 percent CI: 83-93 percent). The additional patient encounters that we detected from clinical text encompassed four National Quality Strategy domains including Communication and Care-Coordination, Patient Safety, Effective Clinical Care, and Person and Caregiver Experience and Outcomes in our assessment. In addition, for five PQRS measures that could not be detected using traditional methods, our enhanced approach made event detection feasible for quality measurement and reporting. Our study suggests that enhanced EHR-based methods have the potential to improve value-based payment arrangements by increasing the detection of clinical events related to physician performance and quality with good precision, by supporting automated reporting from unstructured EHR data across domains that are relevant to multiple stakeholders, including patients, and by expediting the costly and labor-intensive manual chart review process that is now associated with participating in programs such as the PQRS.

References

1. *HHS reaches goal of tying 30 percent of Medicare payments to quality ahead of schedule*, in *A major milestone in the effort to improve quality and pay providers for what works*. 2016: HHS.gov.
2. Casalino, L.P., et al., *US Physician Practices Spend More Than $15.4 Billion Annually To Report Quality Measures*. Health Affairs, 2016. **35**(3): p. 401-406.

3. Fiscella, K., H.R. Burstin, and D.R. Nerenz, *Quality measures and sociodemographic risk factors: to adjust or not to adjust*. JAMA, 2014. **312**(24): p. 2615-6.
4. Lynn, J., A. McKethan, and A.K. Jha, *Value-Based Payments Require Valuing What Matters to Patients*. Jama, 2015. **314**(14): p. 1445.
5. McGlynn, E.A. and E.A. Kerr, *Creating Safe Harbors for Quality Measurement Innovation and Improvement*. JAMA, 2016. **315**(2): p. 129-30.
6. Panzer, R.J., et al., *Increasing demands for quality measurement*. JAMA, 2013. **310**(18): p. 1971-80.
7. Sorrel, A.L., *PQRS Mess*. Tex Med, 2016. **112**(3): p. 59-62.
8. Himmelstein, D.U. and S. Woolhandler, *Physician payment incentives to improve care quality*. JAMA, 2014. **311**(3): p. 304.
9. Zrelak, P.A., et al., *How Accurate is the AHRQ Patient Safety Indicator for Hospital-Acquired Pressure Ulcer in a National Sample of Records?* Journal for Healthcare Quality, 2013: p. n/a-n/a.
10. Zimlich, R., *The PQRS challenge: Will quality metrics improve care or create more reimbursement red tape?* Med Econ, 2013. **90**(7): p. 18-20, 26-8.
11. Yegian, J.M., et al., *Engaged patients will need comparative physician-level quality data and information about their out-of-pocket costs*. Health Aff (Millwood), 2013. **32**(2): p. 328-37.
12. van Walraven, C., C. Bennett, and A.J. Forster, *Administrative database research infrequently used validated diagnostic or procedural codes*. Journal of Clinical Epidemiology, 2011. **64**(10): p. 1054-1059.
13. van Walraven, C. and P. Austin, *Administrative database research has unique characteristics that can risk biased results*. Journal of Clinical Epidemiology, 2012. **65**(2): p. 126-131.
14. Tefera, L., W.G. Lehrman, and P. Conway, *Measurement of the Patient Experience: Clarifying Facts, Myths, and Approaches*. JAMA, 2016.
15. Shanafelt, T.D., et al., *Changes in Burnout and Satisfaction With Work-Life Balance in Physicians and the General US Working Population Between 2011 and 2014*. Mayo Clin Proc, 2015. **90**(12): p. 1600-13.
16. Rosenthal, M.B. and R.A. Dudley, *Pay-for-performance: will the latest payment trend improve care?* JAMA, 2007. **297**(7): p. 740-4.
17. Romano, P.S., et al., *Validity of selected AHRQ patient safety indicators based on VA National Surgical Quality Improvement Program data*. Health Serv Res, 2009. **44**(1): p. 182-204.
18. Manchikanti, L., et al., *Physician Quality Reporting System (PQRS) for Interventional Pain Management Practices: Challenges and Opportunities*. Pain Physician, 2016. **19**(1): p. E15-32.
19. Kern, L.M., et al., *Accuracy of Electronically Reported "Meaningful Use" Clinical Quality Measures*. Annals of Internal Medicine, 2013. **158**(2): p. 77.
20. Horner, R.D., *Risk-adjusted capitation in an era of personalized medicine: a dangerous opportunity to bend the health care cost curve*. Med Care, 2012. **50**(8): p. 633-4.
21. Epstein, R.M., et al., *Measuring patient-centered communication in patient-physician consultations: theoretical and practical issues*. Soc Sci Med, 2005. **61**(7): p. 1516-28.
22. Cassel, C.K. and R. Kronick, *Learning From the Past to Measure the Future*. JAMA, 2015. **314**(9): p. 875-6.
23. *EHR – Based Quality Measurement & Reporting: Critical for Meaningful Use and Health Care Improvement*. 2010, American College of Physicians.
24. Chapman, W.W., *Closing the gap between NLP research and clinical practice*. Methods Inf Med, 2010. **49**(4): p. 317-9.
25. Christianson, J., et al., *What influences the awareness of physician quality information? Implications for Medicare*. Medicare Medicaid Res Rev, 2014. **4**(2).
26. Safran, D.G., et al., *Linking primary care performance to outcomes of care*. J Fam Pract, 1998. **47**(3): p. 213-20.
27. Schultz, E.M., et al., *A systematic review of the care coordination measurement landscape*. BMC Health Serv Res, 2013. **13:** p. 119.
28. Tang, P.C., et al., *Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures*. J Am Med Inform Assoc, 2007. **14**(1): p. 10-5.
29. Iezzoni, L., *Assessing quality using administrative data*. Annals of internal medicine, 1997.
30. Congress, t., *Public Law No: 109-432*, in *H.R.6111 - Tax Relief and Health Care Act of 2006*, t.C. (2005-2006), Editor. 2006, U.S. Government Printing Office.
31. Koltov, M.K. and N.S. Damle, *Health policy basics: physician quality reporting system*. Ann Intern Med, 2014. **161**(5): p. 365-7.
32. Capurro, D., et al., *Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multi-Site Assessment*. eGEMs (Generating Evidence & Methods to improve patient outcomes), 2014. **2**(1).
33. Yetisgen, M., P. Klassen, and P. Tarczy-Hornoch, *Automating Data Abstraction in a Quality Improvement Platform for Surgical and Interventional Procedures*. eGEMs (Generating Evidence & Methods to improve patient outcomes), 2014. **2**(1).
34. Murff, H.J., et al., *Automated identification of postoperative complications within an electronic medical record using natural language processing*. JAMA, 2011. **306**(8): p. 848-55.
35. Matheny, M.E., et al., *Detection of blood culture bacterial contamination using natural language processing*. AMIA Annu Symp Proc, 2009. **2009**: p. 411-5.
36. FitzHenry, F., et al., *Exploring the frontier of electronic health record surveillance: the case of postoperative complications*. Med Care, 2013. **51**(6): p. 509-16.
37. Tamang, S., *CLEVER (CL-inical EV-Ent R-ecognizer)*. 2016, GitHub repository: https://github.com/stamang/CLEVER.
38. Tomas Mikolov, I.S., Kai Chen, Greg Corrado, Jeffery Dean, *Distributed Representations of Words and Phrases and their Compositionality*, in *Neural Information Processing Systems*. 2013.
39. Meystre, S.M., et al., *Extracting information from textual documents in the electronic health record: a review of recent research*. Yearb Med Inform, 2008: p. 128-44.
40. Lowe, H.J., et al., *STRIDE--An integrated standards-based translational research informatics platform*. AMIA Annu Symp Proc, 2009. **2009**: p. 391-5.

41. Services, C.f.M.M. *Measure Codes*. [cited 2017; Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/MeasuresCodes.html.
42. Centers for Medicare & Medicaid Services, *HHS-Operated Risk Adjustment Methodology Meeting*, in *Discussion Paper*. 2016, Centers for Medicare & Medicaid Services, Center for Consumer Information & Insurance Oversight.
43. FANGHUAI HU, Z.S.A.T.R., *Self-Supervised Synonym Extraction from the Web*. JOURNAL OF INFORMATION SCIENCE AND ENGINEERING, 2015. **31**: p. 1133-1148.
44. Wilson Wong, W.L., Mohammed Bennamoun, *Ontology learning from text: A look back and into the future*. ACM Computing Surveys (CSUR) 2012. **44**(4).
45. Michele Banko, M.J.C., Stephen Soderland, Matt Broadhead and Oren Etzioni, *Open Information Extraction from the Web*. Communications of the ACM, 2008. **51**(12).
46. Chapman, B.E., et al., *Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm*. J Biomed Inform, 2011. **44**(5): p. 728-37.
47. Harkema, H., et al., *ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports*. J Biomed Inform, 2009. **42**(5): p. 839-51.
48. Jonnalagadda, S., et al., *Enhancing clinical concept extraction with distributional semantics*. Journal of Biomedical Informatics, 2012. **45**(1): p. 129-140.
49. Hersh WR, P.S., Donohoe L, *Assessing thesaurus-based query expansion using the UMLS Metathesaurus*. Proceedings of the 2000 Annual AMIA Fall Symposium, 2000. **344-348**.
50. Gupta, S., et al., *Induced lexico-syntactic patterns improve information extraction from online medical forums*. J Am Med Inform Assoc, 2014. **21**(5): p. 902-9.
51. Demner-Fushman, D., et al., *UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text*. J Biomed Inform, 2010. **43**(4): p. 587-94.
52. Meng, F. and C. Morioka, *Automating the generation of lexical patterns for processing free text in clinical documents*. J Am Med Inform Assoc, 2015. **22**(5): p. 980-6.
53. Jeffrey Pennington, R.S., Christopher D. Manning, *GloVe: Global Vectors forWord Representation*. Conference on Empirical Methods in Natural Language Processing, 2014.
54. Huang, Y., et al., *Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon*. J Am Med Inform Assoc, 2005. **12**(3): p. 275-85.
55. Yang Xiang, Y.Z., Xiaolong Wang, Yang Qin, and Wenying Han, *Bias Modeling for Distantly Supervised Relation Extraction*. Mathematical Problems in Engineering, 2015. **2015**.

Appendix

Table A1. Physician Quality Measures Developed by the AMA-Convened Physician Consortium for Performance Improvement (PCPI)

NUMBER AND TITLE	MEASURE DESCRIPTION	NUMERATOR STATEMENT	DENOMINATOR STATEMENT
Measure #47 (NQF 0326): *Advanced Care Plan*	Percentage of patients aged 65 years and older who have an advance care plan or surrogate decision maker documented in the medical record or documentation in the medical record that an advance care plan was discussed but the patient did not wish or was not able to name a surrogate decision maker or provide an advance care plan	Patients who have an advance care plan or surrogate decision maker documented in the medical record or documentation in the medical record that an advance care plan was discussed but patient did not wish or was not able to name a surrogate decision maker or provide an advance care plan	All patients aged 65 years and older
Measure #48: Urinary Incontinence Assessment	Percentage of female patients aged 65 years and older who were assessed for the presence or absence of urinary incontinence within 12 months	Patients who were assessed for the presence or absence of urinary incontinence within 12 months	Percentage of female patients aged 65 years and older with a diagnosis of urinary incontinence with a documented plan of care for urinary incontinence at least once within 12 months
Measure #49: Urinary Incontinence Characterization	Percentage of female patients aged 65 years and older with a diagnosis of urinary incontinence whose urinary incontinence was characterized at least once within 12 months	Percentage of female patients aged 65 years and older with a diagnosis of urinary incontinence whose urinary incontinence was characterized at least once within 12 months	Percentage of female patients aged 65 years and older with a diagnosis of urinary incontinence with a documented plan of care for urinary incontinence at least once within 12 months
Measure #191 (NQF 0565): 20/40 or Better Visual Acuity within 90 Days Following Cataract Surgery	Percentage of patients aged 18 years and older with a diagnosis of uncomplicated cataract who had cataract surgery and no significant ocular conditions impacting the visual outcome of surgery and had best-corrected visual acuity of 20/40 or better (distance or near) achieved within 90 days following the cataract surgery	Patients who had best-corrected visual acuity of 20/40 or better (distance or near) achieved within 90 days following cataract surgery	Clinicians who indicate modifier 55, postoperative management only OR modifier 56, preoperative management only, will not qualify for this measure.
Measure #280: Staging of Dementia	Percentage of patients, regardless of age, with a diagnosis of dementia whose severity of dementia was classified as mild, moderate or severe at least once within a 12 month period	Patients whose severity of dementia was classified* as mild, moderate or severe** at least once within a 12 month period	Patient sample criteria for the Dementia Measures Group are all patients regardless of age, with a specific diagnosis of dementia accompanied by a specific patient encounter:

Table A1. Physician Quality Measures Developed by the AMA-Convened Physician Consortium for Performance Improvement (PCPI) (Cont'd)

NUMBER AND TITLE	MEASURE DESCRIPTION	NUMERATOR STATEMENT	DENOMINATOR STATEMENT
Measure #286: Counseling Regarding Safety Concern	Percentage of patients, regardless of age, with a diagnosis of dementia or their caregiver(s) who were counseled or referred for counseling regarding safety concerns within a 12 month period	Patients or their caregiver(s) who were counseled or referred for counseling regarding safety concerns within a 12 month period	Patient sample criteria for the Dementia Measures Group are all patients regardless of age, with a specific diagnosis of dementia accompanied by a specific patient encounter:
Measure #287: Counseling Regarding Risks of Driving	Percentage of patients, regardless of age, with a diagnosis of dementia or their caregiver(s) who were counseled regarding the risks of driving and the alternatives to driving at least once within a 12 month period	Patients or their caregiver(s) who were counseled regarding the risks of driving and the alternatives to driving at least once within a 12 month period	Patient sample criteria for the Dementia Measures Group are all patients regardless of age, with a specific diagnosis of dementia accompanied by a specific patient encounter:
Retired Measure (Jan 1, 2015): *Screening for Dysphagia*	Percentage of patients aged 18 years and older with a diagnosis of ischemic stroke or intracranial hemorrhage who receive any food, fluids or medication by mouth (PO) for whom a dysphagia screening was performed prior to PO intake in accordance with a dysphagia screening tool approved by the institution in which the patient is receiving care	Patients for whom a dysphagia screening was perform ed prior to PO intake in accordance with a dysphagia screening tool approved by the institution in which the patient is receiving care	All patients aged 18 years and older with the diagnosis of ischemic stroke or intracranial hemorrhage who receive any food, fluids or medication by mouth
Proposed Measure: Potentially Avoidable Harmful Events - Urinary Tract Infection	Percentage of patients aged 18 years and older with a diagnosis of ischemic stroke who were hospitalized for seven days or greater, who acquired a Urinary Tract Infection	Patients who acquired a Urinary Tract Infection	All patients aged 18 years and older with a diagnosis of ischemic stroke, who were hospitalized for seven days or greater
Proposed Measure: Potentially Avoidable Harmful Events - Stage III or Greater Decubiti	Percentage of patients aged 18 years and older with a diagnosis of ischemic stroke who were hospitalized for seven days or greater, who developed Stage III or Greater Decubiti	Patients who developed Stage III or Greater Decubiti	All patients aged 18 years and older with a diagnosis of ischemic stroke, who were hospitalized for seven days or greater

CLEVER System Description

Step 1: Terminology Construction

CLEVER combines a *base terminology* and a task-specific *custom terminology* to pre-process the underlying text and populate an event annotation schema that summarizes important information for storage, retrieval and quality measurement event detection. All terms (i.e., character strings that represent words, phrases or symbols) in CLEVER's terminology have a sematic mapping to only one term "class", which groups similar terms together and is used for annotating clinical text with high-level information.

CLEVER's *base terminology* is designed to detect broadly applicable clinical contexts. Each context type is represented as term class. For example, term classes that are associated with context that modify the interpretation of disease condition mentioned in clinical text include negation (e.g., "no evidence of [condition]"), risk (e.g., "risks include [condition]"), screening (e.g., "tested for [condition]"), uncertainty (e.g., "ddx may include [condition]"), temporality (e.g., "past history of [condition]") and familial terms (e.g., "mother passes from [condition]"). Term classes are also used to represent symbols for boundary detection (e.g., "." or ":") polarity (e.g., "+" or "-") and other functional symbols such as the "/" in "no e/o" (short-hand for "no evidence of"). Many of the contexts in the base terminology have been described in prior studies and we include negation and familial terms from the ConText and NegEx systems in CLEVER's base classes.[39,46,47]

Although the base terminology is used with all types of extractors, CLEVER's *custom terminology* is task-specific. Term classes in the custom terminology are called "target" classes, and represent the key clinical concepts relevant to a custom extractor. An example of the target classes, the 2015 and 2016 PQRS study measures for which they are applicable, their class tag and example target terms are shown in A2.

Term expansion methods are popular statistical techniques that have been applied to address some of the methodological challenges associated with constructing comprehensive terminologies for clinical text analysis.[48-50] Based on their co-occurrence with known terms of interest, term expansion methods can be used to detect similar terms without using hand-written rules or labeled sentences.

Biomedical ontologies have been widely applied for clinical information extraction tasks, and provide the backbone for a number of clinical concept recognition systems such as MetaMap, YTEX and the NBCO Annotator; however, unlike biomedical text, which is defined as the kind of unstructured free-text that appears in books, articles, scientific publications and poster, clinical text is written by clinicians in the healthcare setting.[51] For quality measurement event detection, the first challenge posed by adapting biomedical NLP resources to the clinical setting relates to the completeness of terms for tagging a clinical concept.[50,52] A second challenge is presented by the absence of important healthcare concepts from biomedical language resources such as "20/40 or greater vision" for PQRS Measure #191 or "durable power of attorney" for PQRS Measure #47.

Our approach to tailoring our custom terminology for clinical text was to use a *clinical term embedding model* to bootstrap a more complete terminology for quality measurement event detection by learning relevant clinical synonyms from a set of biomedical terms.[38,53] Using over 21 million notes for 1.2 million SHC patients, we identified candidate terms by generating all sequences of up to four contiguous tokens in the

Table A2. Example of Target Classes in our Custom Terminology for PQRS Study Measures

CLASS TAG	CLASS DESCRIPTION	PQRS 2015 AND 2016 PROGRAM MEASURES	EXAMPLE TARGET TERMS
ADVCP	Advanced Care Plan	#47: Advanced Care Plan	*dpoa hc, polst completed, surrogate decision maker, wants aggressive intervention*
DISC	Consultation / Discussion	#47: Advanced Care Plan, #286: Counseling Regarding Safety Concerns, #287: Counseling Regarding Risk of Driving	*counseling regarding, discussed management, consultation with, spent counseling*
UI	Urinary Incontinence (non-specific)	#48: Urinary Incontinence Assessment	urinary incontinence, ui, *leaks urine, wears adult diapers*
UICHAR	UI characterization	#50: Urinary Incontinence Characterization	stress, mixed, *uui, sui*
EYES	20/40 of Better Vision	#191: 20/40 or Better Visual Acuity within 90 Days Following Cataract Surgery	20/20, 20/25, 20/35, 20/40
DSTAGE	Dementia Stage	#280: Staging of Dementia	mild, moderate, severe
SAFE	Safety Concern	#286, #286: Counseling Regarding Safety Concerns,	*safety issues, safety awareness, refrain from operating, safe home environment*
DRIVE	Driving Risks	#287: Counseling Regarding Risk of Driving	*abstain from driving, against driving, refrain form driving, should avoid driving*

Note: Target terms that were learned using term expansion methods appear in italics.

corpus and trained our clinical term embedding model, using a word2vec skip-gram model with one hidden layer.[38] Then, after identifying the set of target classes for each of our study measures with the help of PCPI measure documentation, we identified concept in the UMLS that were relevant to the detection of each measure and used the SPECIALIST system to create a set of initial "seed" terms.[49,51,54] For each seed term, we used our clinical term embedding model to rank and return the top 25 closest (i.e., the most "similar") terms in vector space, or the "nearest neighbors", based on cosine similarity. Examples of terms for each of the target classes used for our 2015-16 PQRS program measures are shown in A2. Terms that were learned using our term expansion techniques appear in italics. Is important to note that all terms, derived from biomedical knowledge-bases or through data-driven term expansion, are manually reviewed to filter spurious or "noisy" terms before they are added to the custom terminology.

Step 2: Pre-processing

Our *pre-processing step* was used to prepare clinical text for downstream analysis and event extraction. For our experiments, we set the scope of candidate event snippets to 125 character to the right and left of a target mention – i.e., a target term documented in clinical text. The character offset length includes white space and symbols to the right and left of a target term.

Figure A1 shows an example of the *intermediate event schema* that is produced by our pre-processing steps for all target mentions detected in clinical text. To enable downstream analysis, our event schema includes structured data, which at minimum is used to define the demographics and combination codes used to determine a patient's eligibility in the denominator a study measure, and the annotated candidate event information from clinical text. In addition to the snippet, and features for indexing and retrieval, note type, and other and term-level information from text, CLEVER provides summary statistics on the top n-grams that occur before and after target terms, and files with n-gram features for each candidate event ID, and the right and left contexts from the snippet that can be used to add more features from other linguistic processing tasks such as concept extraction.

Figure A1. Example of Intermediate Event Schema Template and Annotations for Quality Measurment Event Detection

Template: [candidate event ID*|patient ID|time offset|gender|race|ethnicity|age|note type|target term|term sequence~|class sequence~|snippet|...]

EXAMPLE 1: PQRS Measure #47, Advance Care Plan (*Positive*)

1649714-2|522262|1336|MALE| WHITE|NOT HISPANIC|76|Progress Note, Inpatient|dpoa|DNR DNI- Patient -DPOA-wife|ADDIR-PAT-ADVCP-FAM|Patient is DNR DNI. Patient's DPOA is his wife XXXX. She understands the situation and possible outcome of death.

EXAMPLE 2: PQRS Measure #48, Urinary Incontinence Assessment (*Negative*)

76664|5138|1677299-2|Progress Note, Outpatient|urgency incontinence|no-Denies-urgency incontinence-Denies|NEGEX-NEGEX-UI-NEGEX|MALE|WHITE|NOT HISPANIC|85|no abdominal pain Genitourinary: Denies recent frequency urgency incontinence dysuria and hematuria Musculoskeletal: Denies weakness ...

** Candidate event identifier is unique and assigned by concatenating note id and the target mention number in the note*
~ truncated version of the target term's term and class sequence (Up to two before and one class after) is shown.

Note: The annotation template appears at the top and is summaried for clarity. Targets in the examples are bolded and underlined in orange, and base classes are shown in blue.

Step 3: Extraction

Next we describe the extraction process, which begins with the creation of a rule-based extractor to label a set of development data. For our experiments, we used no more that 1000 randomly selected candidate events to develop and tune each CLEVER extraction rule, which represented less than 1 percent of the total

data for most measures. After a CLEVER rule has been defined, it can be used to label the entire collection of clinical text; or, similar what is done in distantly supervised information extraction methods, it can be used to automatically generate labeled training data for learning a statistical extractor. In the later case, the class sequence that was for the CLEVER rule should be removed before training in order to reduce bias (i.e., building a statistical extractor that "learns-back" the CLEVER rule).[55]

CLEVER Rules

The term and class-level features in the intermediate event schema provide a high-level representation of the underlying text that is intuitive, interpretable, and lends itself to the definition of CLEVER rules. CLEVER rules are intended to use class sequences as input and the main output of the extraction process is a label indicating the *polarity*, or the positive, negative, or neutral status of a clinical event. Unless otherwise specified, rules consider up to two classes before the target mention and up to one class that appears after, referred to as the *truncated class sequence*. For example, using the "#" sign to mark the target class, truncated class sequences that appear in CLEVER's intermediate event schema are shown in Figure 3 and include the "ADDIR_PAT_#ADVCP#_FAM" (advanced directive, patient, advance care plan and family classes) for our positive Advanced Care Plan event and "NEGEX_NEGEX_#UI#_NEGEX" (negation, negation, urinary incontinence and negation classes) for our negative Urinary Incontinence Assessment event.

CLEVER rules have a minimal structure that makes our approach accessible to end users who lack domain and/or linguistic expertise. The two key components of a rule are (1) the inclusion class(es) that must appear in the truncated class sequence and (2) the relevant modifying contexts that are associated with non-positive target mentions. In general, event inclusion rules use target classes in the custom terminology and event exclusion (non-positive) rules use the base classes. Our simplest CLEVER rule was for our Advanced Care Plan measure. Since an event for the measures numerator can include patients that did not have an advance care plan, the class inclusion rule was defined by the advanced care plan target class "ADVCP" for advance care plan and no modifying contexts were considered. However, a positive event for the measure Urinary Incontinence Assessment, was defined by the inclusion of the urinary incontinence class "UI" *and* the absence of the base classes for negation "NEGEX", risk, "RISK", testing and screening, "SCREEN", familial, "FAM", hypothetical, "HYP" and prevention "PREV" terms (see Figure 3 for a negation example). Building on the prior set of Urinary Incontinence Assessment rules, our extractor for Urinary Incontinence Characterization, defined positive events using the subset of terms in the UI class that are specific, i.e., "sui" for stress urinary incontinence; also, by detecting any events with the UI characterization class, "UICHAR" class, which included terms such as "stress" and "urgency", adjacent to the target "UI" class in the truncated class sequence.

In addition to sequence information for terms, CLEVER implements one rule exception that is applicable to all extractors and based on a naïve boundary detection rule. I.e., an exception to the defined earlier is made when the DOT class (indicating a period) is adjacent to the target class on both sides (e.g, NEGEX_DOT_#TARGET#_DOT). This suggests that no modifying terms appear in the sentence that entails our target term; therefore, we always label events with this sequence pattern as positive.

Step 4: Patient-level Reporting

Labels for instances in the intermediate event schema cannot be used on their own for quality reporting; they must be aggregated and further analyzed at the patient-level. To construct a *patient-level candidate event matrix*, we first joined structured and unstructured event features by patient and event; then, for the observation window, we indexed all events by time in days. For quality measurement event detection, we assigned a positive or negative label to each patients; patients with one or more positive events resulted in a positive label and an addition of one to the measure's numerator.