Measuring Care Coordination. Can We Use Data From The Electronic Health Record? Brooke Herndon 1; Timothy P Stablein 2; Carey J. Field 3; Denise L. Anthony 4. 1Dartmouth Medical School and Dartmouth-Hitchcock Medical Center, Norwich, Vermont; 2Dept. of Sociology, Dartmouth College, Hanover, New Hampshire; 3Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire; 4Dept. of Sociology, Institute for Security, Technology, & Society, Dartmouth College, Hanover, New Hampshire. (Proposal ID # 11307)

BACKGROUND: There is widespread agreement that care coordination is a hallmark of high quality health care at all levels. At a national level, the Institute of Medicine recommends focusing on care coordination as a key strategy for improving quality of care for the nation. At a systems-level, researchers have shown that organized, integrated systems of care provide higher quality at lower cost. At a clinical level, numerous professional societies, including the American College of Physicians and the Society of General Internal Medicine, advocate a re-designed primary care environment (The Advanced Medical Home) that emphasizes care coordination. Unfortunately, there is no agreed-upon approach to measuring this critical activity. Due to the resources required, it is not usually feasible to observe large samples of providers or patients directly. However, providers using an electronic health record (EHR) routinely communicate electronically, and this activity is captured and stored digitally. We sought to determine whether the data recording these communications could be retrieved and used to measure intensity of care coordination.

METHODS: We collected data from a convenience sample of 12 primary care physicians (7 family medicine physicians and 5 general internists) working at 3 separate locations (Lebanon, Manchester, and Nashua New Hampshire) within a single health care system in which all providers use the same home-grown EHR. Trained observers (an internal medicine resident and post-doctoral fellow in qualitative research) shadowed each physician for an entire shift and recorded all care coordination activities. Technical support staff from the same institution developed a data query that allowed retrieval of all electronic communications initiated and received by these same physicians on the day that they were observed. We then compared the number of care coordination activities observed directly with the communication events retrieved electronically for each physician and used Spearman’s rho to assess the strength of the association.

RESULTS: A total of 888 care coordination tasks were recorded via direct observation, while the EHR query retrieved a total of 534 communication events. For the 12 individual physicians, the observed counts ranged from 19 to 170 (mean 74), and the electronic counts ranged from 23 to 95 (mean 44). The observed counts were higher than the electronic counts for all but one of the 12 physicians as the electronic counts did not include phone, email, or in-person activities. The difference between the observed and electronic counts ranged from 4 to 49 with a mean difference of 22. Using physicians as the unit of analysis, the Spearman’s correlation between observed and electronic counts was 0.77 (p<0.01). Physicians with the highest observed counts were also more likely to have higher rates of self-initiated electronic communication events (p<0.01).

CONCLUSION: It is difficult to promote and reward care coordination if it can’t be measured. Our findings suggest that EHR data can be used to create a measure of care coordination that is highly correlated with the relative intensity of care coordination measured via direct observation - the gold standard. Additional studies with larger samples and different EHRs are needed to validate these findings.
BACKGROUND: For the healthcare system to achieve its goal of improving quality of care and patient safety, hospital systems and providers must understand which quality improvement (QI) programs are successful. Unfortunately, most QI efforts are analyzed using a pre-post intervention analysis which cannot account for temporal trends nor determine whether improvements are sustained. This abstract outlines a time-series analysis of a Veterans Administration (VA) systems redesign initiative. This initiative, the Flow Improvement Inpatient Initiative (FIX), was a nationwide collaborative with goals of reducing bottlenecks, delays, waste, and errors associated with inpatient care. Two focuses of FIX were to reduce hospital length of stay (LOS) and to increase the proportion of patients discharged before noon. Our objective is to use a time-series analysis to understand the improvements associated with FIX in five outcome measures and whether any improvements are sustained in the two years after implementation.

METHODS: Continuous piecewise linear regression modeling of five risk-adjusted patient outcomes; LOS, in-hospital mortality, 30-day mortality, 30-day all-cause readmission, and rates of patients discharged before noon. Analyzed data covers a 5 fiscal year period (FY05 - FY09) consisting of 1,690,191 discharges from 126 VA facilities. Models were evaluated in SAS version 9.2 using Proc AUTOREG allowing for evaluation of and correction for any autocorrelation between measurements. The analysis focuses on identifying the pre-FIX trend (FY05-06) for each outcome measure and then identifying whether there was any change during FIX (FY07) and if the change was sustained after FIX (FY08-09).

RESULTS: All five outcome measures show a steady rate of improvement throughout the 5-year study period. The modeled rate of change per year for each outcome is displayed in Table 1. Of the outcome measures, LOS, 30-day mortality and discharge before noon show improvements in FY07 that are in the desired direction and statistically different than the change predicted by the baseline rate. Both LOS and 30-day mortality show continued improvements throughout FY08 and FY09. Conversely, discharges before noon shows a leveling in FY08 and a decline in FY09. In-hospital mortality and 30-day all-cause readmission show significant changes in the undesired direction during FY07.

CONCLUSION: These results show that during FIX improvements beyond baseline temporal trends occur in three of the outcome measures, with two of them showing signs of sustainability. In contrast, if FIX was analyzed using a pre-post study design, likely all five outcome measures would exhibit improvements during FIX. Most importantly this analysis highlights the difficulties of quality improvement in healthcare by showing that achieved results are not always successfully sustained. It appears that VA’s continued focus on reducing LOS has led to not only sustaining the results of FIX but the achievement of additional improvements. However, there has been less long term focus on discharges before noon and it appears those results have not been sustained. Lastly, this time-series analysis shows that FIX did not lead to improvements in outcomes beyond its scope, with in-hospital mortality and 30-day all-cause readmission rates remaining level during and after FIX. This time-series analysis shows FIX had initial success in meeting its objectives, but there is still much to be achieved in the process of obtaining and sustaining optimal patient flow in VA.
Use of Close Follow-Up as a Strategy to Mitigate Harm from Diagnostic Error in Primary Care

Hardeep Singh; Traber Davis Giardina; Samuel Forjuoh; Michael Reis; Steven Kosmach; Myrna Khan; Eric Thomas.

1VA Health Services Research and Development Service (VA HSR&D) Center of Excellence, Michael E. DeBa, Houston, Texas; 2Michael E. DeBakey Veterans Affairs Medical Center and Baylor College of Medicine, Houston, Texas; 3Scott & White Healthcare System and A&M Health Science Center, Temple, Texas; 4University of Texas at Houston & Memorial Herman Center for Healthcare Quality and Safety, Houston, Texas. (Proposal ID # 11426)

BACKGROUND: Diagnostic errors in primary care are harmful but not well studied. Ensuring close follow-up might be one strategy to minimize harm from diagnostic error, but little empirical work has examined diagnostic error and the quality of follow-up in primary care. Because visits with errors might involve greater provider uncertainty, we hypothesized that follow-up practices (e.g., scheduling follow-up, scheduling follow-up at an appropriate time interval) by providers would differ in situations involving diagnostic errors versus those with no diagnostic errors.

METHODS: We applied electronic trigger queries to electronic health record (EHR) repositories at two large health systems between October 1, 2006 and September 30, 2007 to identify records likely to contain a diagnostic error. One site was an urban VA facility with 5 on-site and 5 satellite clinics, the other a large private health care system with 4 community-based clinics. Both sites provided longitudinal care in relatively closed systems and had integrated and well-established EHRs. Trigger queries were 1) primary care index visits followed by unplanned hospitalization within 14 days, and 2) primary care index visits followed by ≥ 1 unscheduled visit(s) within 14 days. Two physicians independently reviewed cases meeting either criterion for evidence of diagnostic errors and follow-up practices. For each index visit, reviewers recorded whether the provider scheduled any future follow-up visit, and if so, whether the provider scheduled routine follow-up (i.e., non-urgent, next available follow-up for routine medical issues usually in several months) or close follow-up (i.e., within 7 or 30 days depending on the clinical situation and specifically given for uncertain diagnosis or close monitoring). Reviewers also rated whether the time interval between the index visit and follow-up was judged suboptimal (i.e. not close enough) in a higher proportion of index visits with errors, although this was not statistically different from index visits without errors (36.6% v. 9.3%, \( P = .07 \)).

RESULTS: In 212,165 visits, we found diagnostic errors in 177 of 997 (17.8%) triggered records. Errors included a large variety of clinical conditions, and no single condition accounted for more than 10% of all errors. Index visits with and without diagnostic errors were equally likely to have had any future follow-up scheduled by the provider (57.0% v. 60.9%, \( P = .34 \)). Similarly, comparable proportions of index visits with and without diagnostic errors were scheduled for future "routine" follow-up (rather than close follow-up) at the time of the visit (43.6% v. 42.3%, \( P = .81 \)). The time interval between the index visit and follow-up was judged suboptimal (i.e. not close enough) in a higher proportion of index visits with errors, although this was not statistically different from index visits without errors (36.6% v. 9.3%, \( P = .07 \)).

CONCLUSION: There were no differences in follow-up practices between primary care visits with and without diagnostic errors, suggesting that close follow-up might be underutilized to mitigate harm in certain visits with diagnostic error. Our findings suggest that providers might not always perceive greater clinical uncertainty or the need to initiate close monitoring in situations involving diagnostic errors. Future investigation on describing and utilizing appropriate follow-up practices in primary care might prevent patient harm from diagnostic error.
General Medical Versus Geriatric Co-Morbidity Counts: Opposite Effects on Overall Quality of Care in Complex Geriatric Primary Care Patients

BACKGROUND: Prior research across multiple data sources suggests that patients with greater burden of co-morbid conditions receive better -- rather than worse -- quality of care. Because the treatment of geriatric conditions is time-consuming and often falls outside the traditional medical model, we hypothesized that the total burden of geriatric conditions would be associated with lower overall quality. Using data from the Assessing the Care of Vulnerable Elders-2 (ACOVE-2) study, a study that focused on the care of patients with both geriatric and medical conditions, we assessed the association between overall quality of care with both the number of general medical conditions and the number of geriatric conditions.

METHODS: As part of the ACOVE-2 study, 644 patients age >=75 who screened positive for symptoms of at least one of 3 geriatric conditions were enrolled in a practice-based intervention to improve the care of dementia, falls, and urinary incontinence. To assess care quality, we constructed an overall quality of care score that was a composite of 98 process of care quality indicators (QIs) that measured the care of preventive, general medical, and geriatric conditions (# of QIs passed divided by number of QIs eligible) over a 13 month observation window for each patient. We also constructed separate counts of general medical conditions (coronary artery disease, atrial fibrillation, congestive heart failure, cerebrovascular disease, diabetes, and chronic obstructive lung disease) and geriatric conditions (dementia, falls or fear of falling, bothersome urinary incontinence, osteoporosis, hearing impairment, malnutrition) documented in the medical record during the observation window for each patient. To assess association between co-morbidity counts and overall quality we used multivariable regression, controlling for age, gender, vulnerability to death and decline, number of primary care visits, and ACOVE-2 control vs intervention site.

RESULTS: The mean number of general medical conditions was 1.9 (SD 1.3, range 0-6) and the mean number of geriatric conditions was 1.6 (SD 0.8, range 0-4). The two counts were uncorrelated (r=.04, p=.3). Nearly all (95%) had at least one conditions in both categories, and more than half (52%) had at least 2 conditions in each category. On average, each additional general medical condition was associated with a 5% point increment in overall quality, while each additional geriatric condition was associated with a 3.2% point decrement, independent of each other and the multivariable controls (p<.001 for both). A moderately-complex patient with 1 general medical and 1 geriatric co-morbidity had a predicted overall quality of 53% (95% CI 50-57%). Adding 2 additional general medical co-morbidities to this hypothetical patient increased the predicted overall quality to 60% (95% CI 57-64%) whereas adding 2 additional geriatric co-morbidities decreased expected overall quality to 42% (95% CI 38-46%).

CONCLUSION: While a greater number of general medical conditions was related to better quality of care, patients with a greater burden of geriatric conditions received worse overall quality of care, suggesting a need to focus on improving care for patients with multiple geriatric conditions.
**Does Public Reporting Impact Quality Of Care In Wisconsin?** Geoffrey C Lamb 1; Maureen Smith 2; William B Weeks 3; Alexandra Wright 2; Daniel Gottlieb 3; Matthew Gigot 4; Lucy Stewart 1. 1Medical College of Wisconsin, Milwaukee, Wisconsin; 2Health Innovation Program, University of Wisconsin, Madison, Wisconsin; 3The Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, New Hampshire; 4Wisconsin Collaborative for HealthCare Quality, Madison, Wisconsin. (Proposal ID # 12280)

**BACKGROUND:** The Wisconsin Collaborative for Healthcare Quality (WCHQ) is a voluntary, statewide consortium of physician groups, hospitals, health plans and employers working together to improve health care in Wisconsin by publicly reporting comparative measures of healthcare quality. However, there has been no formal evaluation of the impact of the WCHQ's public reporting efforts. This project was designed to establish whether public reporting of ambulatory quality measures by WCHQ is associated with improvement in the delivery of recommended interventions and outcomes. To separate the effects of public reporting from other trends, the project took a three-pronged approach: 1) determine whether there was improvement among WCHQ participants with respect to the measures being reported; 2) survey participants to see how they responded to the information as it was reported; and 3) compare the rate of improvement within the WCHQ to areas not participating in the WCHQ.

**METHODS:** WCHQ member groups commit to report outcomes for 13 ambulatory quality measures annually. They collect and report their own data either from random chart review, electronic capture, or a combination of the two (hybrid). All results are audited using a standardized method. **WCHQ longitudinal analysis** - In this analysis each measure was assessed to determine if there was an improvement in the mean performance of all organizations. A pairwise t-test and Tukey's range test were performed on each measure for each year. Subsequent analysis was performed to determine how many years were required to achieve statistically significant improvement for each measure. **Survey** - The University of Wisconsin Survey Center conducted a mail survey of the physician groups and their related clinics. The survey contained 3 sections: clinic characteristics; whether projects were undertaken in response to WCHQ reporting; and specific improvement initiatives. **Comparisons to non-WCHQ participants** - The Dartmouth Institute (TDI) worked with MMI, a subsidiary of IMS Health. The MMI dataset allows for identification of physicians who work at specific sites. Using a 20% sample of Medicare beneficiaries, TDI assigned patients to physicians based on a majority of their visits, giving priority to primary care MDs. Based on these assignments, patients were characterized as WCHQ related, Wisconsin-Non WCHQ, Iowa/South Dakota and the rest of the United States. 3 diabetes process measures were captured: eye exams, lipid profiles and HgbA1c tests. Mammography was determined for women aged 67-69. Compliance rates were calculated and compared between sites and for each year 2004 - 2007.

**RESULTS:** For WCHQ as a whole, each measure showed an improvement in performance during the study period 2004 - 2008. This improvement was statistically significant in all measures reported for at least 3 years (HgbA1c testing, LDL testing, LDL control, kidney function screening, BP control, breast Ca screening and colorectal Ca screening). The survey of physician groups found that it was very common for WCHQ member organizations to formally focus on WCHQ measures during the study period. 15 of 18 groups reported giving priority to at least one WCHQ measure in response to WCHQ reporting. 9 groups indicated that their priorities were always or nearly always in response to WCHQ reporting. 6 showed a mix of responses, with 5 of those only occasionally responding to WCHQ reporting. Looking at the Dartmouth measures, WCHQ participants consistently outperformed the comparator groups in measures that are publicly reported through the Collaborative (HgbA1c testing, lipid testing in diabetics and breast cancer screening). In each of these measures both the overall performance and the rate of improvement during the study years was higher for WCHQ participants. Of note the only measure in which WCHQ participants failed to perform as well as one of their comparison groups was in diabetic eye testing which is not publicly reported by the Collaborative.

**CONCLUSION:** The three components of this study provide compelling evidence that public reporting of ambulatory measures led to improved performance among WCHQ participants. Over the time frame that public reporting was in place overall performance of the group improved significantly. Participants when surveyed stated that they focused improvement efforts in response to their performance on reported measures. Most significantly, although performance on many of these measures improved elsewhere, the members of the WCHQ improved at a faster rate than national comparison groups on measures reported by WCHQ.

A recent systematic review of public reporting concluded that "rigorous evaluation of many public reporting systems was lacking", and what evidence of effectiveness does exist is largely hospital based. This study helps to address this gap, especially in the ambulatory environment. It provides strong support for the concept that public reporting really can lead to improved performance.
Development and Validation of a Natural Language Processing Computer Program to Measure the Quality of Colonoscopy Ateev Mehrotra; Hendrik Harkema. University of Pittsburgh, Pittsburgh, Pennsylvania. (Proposal ID # 12447)

BACKGROUND: Quality measurement has been hampered by the costs and burden of reviewing medical charts and the limited information available in medical claims. The implementation of Electronic Health Records (EHRs) is expected to facilitate quality measurement; however, much of what is captured in EHRs remains in "free text", and thus still requires manual abstraction of the information. Natural language processing (NLP) is a field of computer science in which relevant structured data is abstracted from free-text language. There is great interest in whether NLP applications can be developed to extract relevant information from EHRs for quality measurement. In this project, we developed and validated a NLP-based computer program for measuring the quality of colonoscopy and associated pathology reports. Applying NLP to the assessment of colonoscopy is an ideal place to apply a NLP tool because colonoscopy reports are typically in electronic format and poor quality of colonoscopy has been associated with higher incidence of colorectal cancer.

METHODS: A NLP computer program was developed that abstracts the necessary data to measure published colonoscopy quality indicators from major gastroenterology societies, including documentation of cecal landmarks and bowel preparation quality, reporting withdrawal time, and making appropriate follow-up recommendations. The NLP tool was then tested on a new validation set of 453 colonoscopy and 226 associated pathology reports that were also manually reviewed by 3 physicians. Reports were randomly sampled from the 32,485 colonoscopy reports from the 10 hospitals in the University of Pittsburgh Medical Care system in 2008-9.

RESULTS: Overall performance on most quality indicators, as measured by manual review of charts, was poor. For example, adequate withdrawal time was documented in only 0.8% of reports and appropriate follow-up recommendations in only 24%. Compared to the manual review, the NLP tool had a mean accuracy of 86.7 (SD 10.7). The difference between the clinician-abstracted and NLP-abstracted quality scores varied from 0.2 to 13.6%.

CONCLUSION: This program represents one of the first NLP based quality measurement tools, but to be used widely the program needs further refinement and validation. The disagreement on certain measures between manual abstraction and the NLP tool helps to highlight key limitations of NLP-based quality measurement applications. As the use of EHRs grows, there is great potential for NLP-based programs to automatically assess the quality of care. Our project highlights key strengths and limitations of this approach.