Automating Performance Measures and Clinical Practice Guidelines: Differences and Complementarities

[1]Samson W. Tu MS, [2]Susana Martins MD MSc, [2]Connie Oshiro PhD, [2]Kaeli Yuen, [2]Dan Wang PhD, [2]Amy Robinson PharmD, [2]Michael Ashcraft MD, [1,2]Paul A. Heidenreich MD MS, [1, 2]Mary K. Goldstein MD MS

[1]Stanford University, Stanford, CA; [2]VA Palo Alto Health Care System, Palo Alto, CA

Abstract

Through close analysis of two pairs of systems that implement the automated evaluation of performance measures (PMs) and guideline-based clinical decision support (CDS), we contrast differences in their knowledge encoding and necessary changes to a CDS system that provides management recommendations for patients failing performance measures. We trace the sources of differences to the implementation environments and goals of PMs and CDS.

Introduction

Performance measures and clinical decision support (CDS) are methods to improve quality of care. Both performance measures and CDS systems rely on clinical evidence, often summarized in clinical practice guidelines (CPGs), to define the standards of care. They are inter-related but distinct. Performance measures seek to improve care by retrospectively measuring the quality of the care provided to populations of patients, while CDS focuses on prospectively providing evidence-based therapeutic recommendations and alerts that are custom-tailored for the circumstances of particular patients. CDS may include performance measurement information as feedback to health professionals; however, in this paper we will use the term CDS in the sense of providing timely information and advisories to health professionals to assist decision-making. Once performance measures have been established, health care systems can provide CDS in order to improve performance on the items that are being measured. To improve the quality of care beyond providing aggregated data, the first step is feedback about performance for each patient with respect to the targets being measured. The next step is to add CDS to give additional recommendations about how to manage the condition to achieve the target

In this paper we seek to concretely characterize the ways performance measures and guideline-based CDS differ yet can be complementary to each other by closely analyzing two pairs of performance-measure and CDS implementations. In the first pair (Analysis 1), we examine the implementations of similar clinical recommendations as performance measures and as CDS, and highlight how the implementations differ in their workflow integration, cohort definitions, definition of compliance, use of data, and output formats. We categorize the rationales for the divergence in inclusion and exclusion criteria of performance measure and CDS. In the second pair (Analysis 2), where a CDS system is used to provide guidance on the management of patients who have failed particular performance measure, we describe the necessary changes to the implementation of CPGs in a CDS system and how the recommendations of the CDS system can complement performance measure status in a provider's dashboard.

Background

The Veteran Health Administration (VHA) of the Department of Veterans Affairs (VA) has been a leader in health care quality assessment and improvement [1]. Not only did it implement quality-improvement efforts that were guided by performance measurements [2], it also pioneered the application of evidence-based CPGs at points of care [3]. For more than ten years, our group at VHA Palo Alto Healthcare System has used the ATHENA CDS system, a knowledge-based system to provide CDS for guideline-based care, to investigate issues related to the operationalization, testing, and deployment of CPGs [4-7]. The basic system architecture includes Protégé [8] knowledge bases (KBs) that contain computer-interpretable CPG recommendations encoded using a domain-independent guideline model, a guideline interpreter execution engine that applies the encoded recommendations to patient data to generate patient-specific recommendations, and client programs that display the recommendations to and interact with CDS users. Initially focused on hypertension, the ATHENA CDS system now includes knowledge bases in several other clinical domains, such as hyperlipidemia, chronic kidney disease (CKD), diabetes mellitus (DM), heart failure (HF), and opioid therapy for non-cancer chronic pain [9-11].

For each knowledge base, the ATHENA CDS system evaluates decision criteria to determine guideline-concordant management goals and recommended actions for a patient. This computational infrastructure of the CDS system

would seem well suited to evaluate performance measures as well, since implementing performance measures involves using similar data and criteria. Performance measures focus on numerator and denominator criteria to determine whether a patient is included in the target population of the performance measure (the denominator criteria), and, if included, whether their care satisfies the definition of quality care (the numerator criteria). In 2011, our group was afforded the opportunity to study the processes and results of automating performance measures and guideline-based recommendation for patients diagnosed with heart failure in ATHENA CDS.

In a subsequent effort, we adapted the ATHENA CDS system to provide guideline-based recommendations to improve the care of patients who fail performance measures. Within the VA, the Veterans Integrated Service Network (VISN) 21 Pharmacy Benefits Management (PBM) group has developed a clinical data warehouse, based on an SQL Server database. The PBM group has built a clinical dashboard for use by both managers and individual providers, including nurses, pharmacists, physicians, and other members of the health care teams. The dashboard provides tools to monitor the clinical performance measures used by VA, focusing on the diabetes, heart disease, and hypertension measures as the priority areas identified by the leadership. The clinical dashboard is available to clinical managers for information about the primary care providers they manage, and to primary care providers for managing their own panels of patients. The dashboard provides a stoplight-type report (red/yellow/green on each performance measure) which can be viewed both as a "panel" view (a provider's panel of patients) or as a visit view (patients coming into clinic today, for visit planning). Our group undertook a project to complement the dashboard's implementation of performance measures with detailed CDS for patients who fail to satisfy these performance measures. This opportunity allows us to examine what is different about a CDS system used in the electronic health record vs a CDS system used within a performance measure dashboard.

Method

To investigate concretely how automated performance measure systems and guideline-based CDS systems differ from and complement each other, we performed two analyses.

Analysis 1: Comparing Implementations of Heart Failure Performance Measure and CDS Systems

The analysis consists of comparing a heart-failure performance-measure (HF-PM) system that implements National Qualify Forum (NQF) Measure 0081 on the use of angiotensin-converting enzyme (ACE) inhibitor or angiotensin receptor blocker (ARB) therapy for left ventricular systolic dysfunction [12][1] and the ATHENA HF-CDS system that implements similar recommendations from the 2013 ACC/AHA guideline for the management of heart failure [13] (See Figure 1).[2]

ATHENA CDS systems structure recommendations in terms of the EON guideline model [14, 15]. This guideline model formalizes a CPG as a knowledge structure containing eligibility criteria, goals or targets of therapeutic interventions, and a clinical algorithm that provides distinct decisions and action choices for patients in various clinical scenarios. The guideline model includes expression languages for performing queries and for encoding decision criteria [14]. At run-time an expression-evaluation module of the execution engine uses patient data to evaluate expressions and conclude whether a decision criterion evaluates to true or false for a patient. The evaluation of decision criteria helps to generate therapeutic recommendations appropriate for a particular patient.

To implement performance measures, we extended the EON modeling and execution infrastructure. Because we wish to compute a collection of performance measures using the same data set, we organize performance measures into groups, such as measures applicable to inpatient cases and measures that are applicable to outpatient cases. To improve system efficiency, we identified, for each group, common criteria that are applicable to all measures in the group. (For example, for the outpatient performance measures, one required criterion is that a patient has an outpatient encounter during the measurement period.) These common criteria (implemented as EON eligibility

[1] We actually implemented NQF performance measures 0081 and 0083, where the NQF 0083 is a measure that evaluates the use of beta blockers for patients with heart failure. For the sake of simplicity, we report the results derived from the use of ACE inhibitor and ARB. The conclusions that can be drawn from the use of beta blocker are similar.

[2] Specifically, we implemented recommendations related to ACE inhibitors and ARBs in [13]: **'**ACE inhibitors are recommended in patients with HF*r*EF and current or prior symptoms, unless contraindicated, to reduce morbidity and mortality. (*Level of Evidence: A*)' – Yancy 7.3.2.2.' and **'**ARBs are recommended in patients with HFrEF with current or prior symptoms who are ACE inhibitor intolerant, unless contraindicated, to reduce morbidity and mortality (*Level of Evidence:A*) – Yancy 7.3.2.3

criteria) are part of a performance measure's denominator criteria. Each performance measure within a group has a set of inclusion criteria (all of which must evaluate true for a patient to be in the denominator population), a set of exclusion criteria (any of which, if evaluated to true, excludes a patient from the denominator population), and a set of criteria to achieve (any of which, if evaluated to true, puts the patient in the numerator population). Because modeling a performance measure's numerator and denominator criteria uses the well-tested EON expression language for encoding the criteria and executing them against patient data, these extensions were easily implemented.

While NQF 0081 provides an initial level of specifications regarding the numerator, denominator inclusions, and denominator exclusions for these measures, in order to operationalize the computation of the measures we had to interpret the measures in much more detail. For instance, the denominator exclusions in NQF #0081 are defined quite broadly as follows: "Documentation of medical reason(s) for not prescribing ACE inhibitor or ARB therapy", "Documentation of patient reason(s) for not prescribing ACE inhibitor or ARB therapy", "Documentation of system reason(s) for not prescribing ACE inhibitor or ARB therapy." In order to better specify such broad exclusion criteria, we consulted other heart failure performance measures that had more specific definitions for denominator exclusions. For instance, the VA External Peer Review Program (EPRP) has a FY2012 technical manual that specifically defines an inpatient performance measure of "HF patients with left ventricular systolic dysfunction (LVSD) who are prescribed an ACE INHIBITOR or ARB at hospital discharge." The measure's denominator exclusions include: "Patients who had a left ventricular assistive device (LVAD) or heart transplant procedure during hospital stay" and "Patients who have a Length of Stay greater than 120 days". Similarly, other NQF-endorsed heart failure performance measures related to ACE inhibitors and ARBs, such as NQF #0610, specify additional denominator exclusions, such as "Evidence of metastatic disease or active treatment of malignancy (chemotherapy or radiation therapy) in the last 6 months." We have pooled the exclusion criteria from these various sources and used them as operationalized denominator exclusions of NQF #0081.

We evaluated the performance of HF-PM using a convenience sample of 340 VA patients. Out of the 340 patients, 73 outpatient cases and 33 hospitalizations satisfy initial eligibility criteria. A preliminary validation of the accuracy of the system on 12 inpatient hospitalizations and 20 outpatient cases demonstrates that the system successfully generates conclusions for the ACE-inhibitors/ARB and beta-blockers performance measures in the majority of cases

To operationalize the ACE inhibitor and ARB recommendations in the 2013 ACC/AHA guideline, we followed the methodology outlined by Shiffman [16] for making explicit the translation of document-based knowledge: markup of the text; atomization; de-abstraction; disambiguation of concepts; verification of completeness; and addition of explanations. We have identified additional steps not in Shiffman's categorization and have reported them in a separate paper [17]. In the rest of the paper, we refer to the ATHENA CDS system that implements the heart-failure CPG recommendations as HF-CDS.

Implementing similar recommendations, first as performance measures and then as part of a CDS system, afforded us the opportunity to systematically analyze how different usages of the same recommendations have implications for workflow integration, cohort definition, definition of compliance, use of data, and output formats. We describe the findings in the Results section.

Analysis 2: Use Guideline-Based CDS to Complement Performance Measure Evaluation

The second analysis involves a production implementation of performance measures in the VA VISN 21 dashboard and a modified ATHENA CDS system (called ATHENA PMtoCDS) designed to provide decision support on the management of patients who have failed the performance measures. To provide recommendations that are consistent with performance measure evaluations, we made the necessary changes to the existing CDS implementations of guidelines for the management of type II diabetes mellitus, hyperlipidemia, heart failure, chronic kidney failure, and hypertension. We also developed a prototype user interface that integrates the outputs of the performance-measure dashboard and ATHENA PMtoCDS. The user interface allows a user to drill down from the top-level display of performance measure evaluations to see recommendations from ATHENA PMtoCDS on how to improve compliance with the performance measures (Figure 2). In this pairing of performance-measure and CDS systems, we demonstrate the complementarity of the two systems in achieving the institution's clinical objectives.

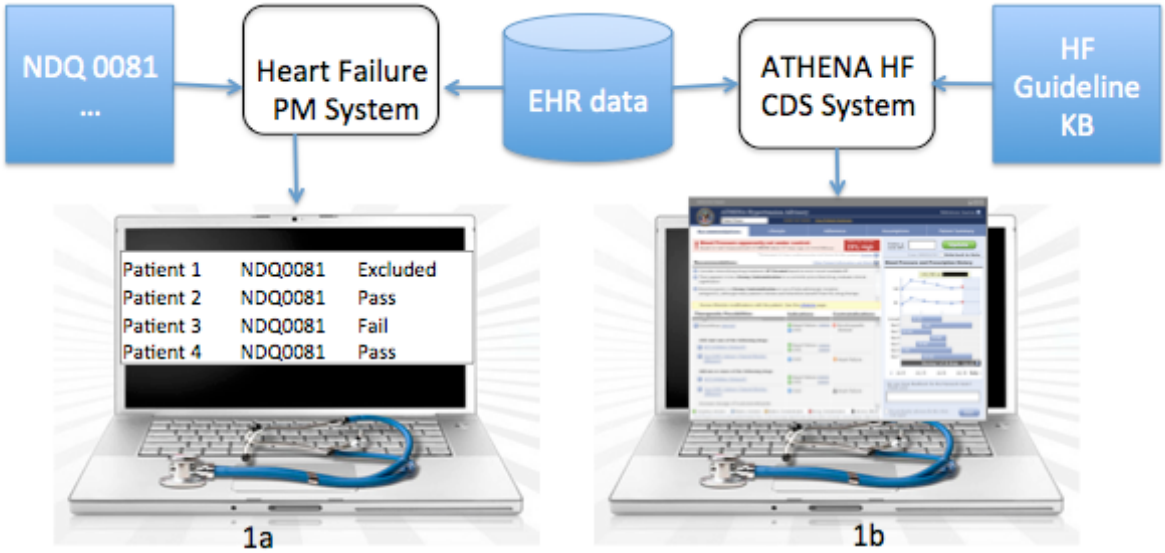

Figure 1. In the first pair of systems, the Heart Failure Performance Measure system (1a) is designed to generate reports on whether a cohort of patients satisfies the NQF 0081 performance measure. ATHENA HF CDS System (1b) is designed to generate detailed management recommendations for guideline-based care at the time of a patient encounter

To make the CDS advisories consistent with performance measures, we modified the original encodings of the guideline recommendations. The advisories generated by ATHENA PMtoCDS for patients who fail specific performance measures are passed to the dashboard for display to users in the context of the performance measure evaluations. Parallel to Analysis 1, we report as results our analysis of the workflow integration, cohort definition, compliance definition, use of data, and design features of user interface of the CDS. In this analysis our focus is on changes that are necessary to integrate CDS recommendations in the context of a performance-measure–oriented dashboard.

Results

Table 1 at the end of section highlights the main results of the analyses.

Analysis 1: Comparing Implementations of Heart Failure Performance Measure and CDS Systems

We will contrast performance measures and guideline-based CDS implementations in terms of workflow integration, cohort definitions, compliance definition, use of data, and output structure.

1. Workflow Integration

Both the HF-PM and the HF-CDS systems are research prototypes that were not deployed in actual clinics. The design of the HF-PM system involves the system processing, in batch mode, the data of a cohort of patients to see whether their treatments satisfy performance measures. The HF-PM system evaluates performance measures for both inpatient and outpatient cases. Both aggregated results and results for individual patients are stored in a database accessible to providers or administrators when desired. The HF-CDS system, on the other hand, evaluates patient-specific current clinical care and prospectively recommends best practices based on the encoded guideline recommendations. The CDS system also brings relevant patient data into one display with layered information to reduce the cognitive burden of searching the EHR for the information required for decision-making. Past ATHENA CDS deployments focused on providing decision support to primary care providers in outpatient clinics. For example, ATHENA HTN, the hypertension management version of ATHENA CDS, would pop up a CDS window when a provider selected a patient who was eligible for guideline-based care. The window would contain recommendations based on the JNC/VA guideline for the management of hypertension and the available patient data [7].

2. Cohort Definitions

Next we report the results of the comparison between HF-PM and HF-CDS systems in terms of identifying patients who should be included in performance measure evaluations and CDS support. As described in the Methods section,

we use as a case study the ACE inhibitor recommendation based on NQF 0081 [12] and the 2013 HF guidelines [13]. To make the criteria comparable, we examine the cohort definition of the outpatient component of NQF 0081 only.

HF-CDS makes use of 13 criteria that determine whether or not a recommendation should be made on ACE inhibitor or angiotensin II receptor blocker (ARB) for a patient with heart failure. HF-PM uses 33 criteria to select the cohort of patients eligible to have a prescription for an ACE inhibitor or an ARB.

We identified 8 criteria that are identical, 16 HF-PM criteria that are handled differently in HF-CDS, 10 criteria that are in HF-PM only and 3 criteria that are in HF-CDS only.[3]

- Similar criteria differed in definition and modeling choices

  We organize the 16 instances in which similar criteria were handled differently in HF-PM and HF-CDS systems into 3 categories:

  a. Role of clinical judgment: HF-CDS alerts providers to conditions that require clinical judgment when recommending ACE inhibitor while HF-PM excludes these patients from the cohort/denominator in order to improve specificity. Performance measures exclude specific reported adverse events such as hypotension, hyperkalemia, worsening renal function due to ACE inhibitor or ARB. HF-CDS displays both the recommended drugs and the adverse events and allergies to providers and leaves the choice of whether to prescribe the recommended drugs to their judgment. Other clinical conditions were also excluded by performance measure, such as presence of aortic stenosis, hypertrophic cardiomyopathy, renal artery stenosis, stage 3 chronic kidney disease, eGFR between 30 and 59, and active prescription for aliskiren, are handled in the HF-CDS by explicitly alerting the provider to these conditions.
  b. Differences in sources: First, the heart-failure guideline used as the basis for HF-CDS does not contain an enumeration of ICD 9 codes used to define heart failure. Subject matter experts weighed in to define the list of ICD 9 codes and included cardiomyopathy codes that were excluded in the sources that we used to define heart failure for NQF 0081. Second, the definition of the thresholds for ejection fraction in the performance measure was more stringent (less than 40) than the threshold used for HF-CDS (less than or equal to 40).
  c. Differences in the timing and retrospective/prospective nature of performance measure and CDS: Pregnancy was modeled differently in the two systems. In HF-PM we look at the data retrospectively for pregnancy codes in order to exclude pregnant women from the denominator population. HF-CDS was designed to give close to real time advice. Given that many VA patients receive pregnancy care outside of VA and thus the relevant pregnancy codes may not be up-to-date in VA data, we decided to issue an alert to all women of childbearing age about the use of ACE inhibitor.

- Criteria unique to HF-PM

  HF-PM had 10 unique criteria that are not applied in HF-CDS. They can be grouped into two categories:

  a. Active in health care system: HF-PM excludes patients who are not active in the health care system using criteria such as visit in past 12 months and absence of death. In HF-CDS there is no need for this filter since CDS applies to patients with a scheduled visit, and are already known to be active in the health care system.
  b. Differences in sources: Our modeling of performance measure was based on documentation from national sources that mentioned exclusion criteria that were not cited in HF guidelines. These include use of hydralazine prior to ACE inhibitor/ARB (NQF 0610), multiple myeloma (NQF 0610), active prescription for pulmonary hypertension medications (NQF 0610), heart valve surgery (NQF 0610) and previous admission for hyperkalemia (NQF610).
  c. Differences in goals: HF-PM excludes patients with potential limited life expectancy while the HF-CDS offers the recommendations since the life-prolonging measures are often also measures that reduce symptoms and contribute to quality of life.

[3] The categorized HF-PM criteria sum to 34 because one HF-PM criterion, eGFR<60, is broken up into eGFR<30 and 30<=eGFR<60, as HF-CDS uses them differently. The categorized HF-CDS criteria do not sum to 13 because multiple HF-PM criteria (e.g., adverse reactions) are handled uniformly in HF-CDS (e.g., adverse reactions not treated as absolute contraindications are displayed with recommended drugs).

- Criteria unique to HF-CDS

  The three criteria that are unique to HF-CDS-only can be classified into three types:

  a. Differences in sources: HF-CDS used creatinine values in addition to eGFR and ICD 9 codes for chronic kidney disease to exclude patients.
  b. Difference in the scope of CDS: Because of the complexity of managing end-stage heart failure, stage D patients are excluded from the scope of HF-CDS.
  c. Differences in output: Because HF-CDS generates recommendations on the choice of medications to prescribe, the presence of ACE inhibitor or ARB on the current medication list suppresses such recommendations.

3. Definition of Compliance

The NQF 081 performance measure defines a patient satisfying the denominator criteria as compliant with the performance measure if the patient was prescribed either ACE inhibitor or ARB. HF-CDS, on the other hand, preferentially recommends ACE inhibitor as the drug of choice, because of its well-founded evidence base and lower cost, and recommends ARB only if a patient is intolerant of ACE inhibitor.

4. Use of Data

Although both systems have access to all data in the EHR we note a distinction in temporal requirements when defining eligible patients. ATHENA CDS uses all data available in the electronic medical record with limited temporal restrictions on specific criteria while performance measure applies stricter temporal limits in many inclusion and exclusion criteria. For example, the NQF 081 performance measure requires the presence of a heart failure ICD9 code in the 2 years prior to the measurement period while for HF-CDS it is sufficient to have an ICD9 code for heart failure at any time. This difference supports the goals of the distinct systems: in performance measure to improve the specificity of the cohort leading to better credibility of results and in CDS to provide full presentation for review by providers in the context of care.

5. Output Formats

The performance measure primary output is geared to administrators wishing to evaluate the quality of care provided to patients at a given site or over a specific region or to health care professionals to identify missed opportunities to improve care. For HF performance measure, the raw output for each patient consists of a listing of each inclusion, exclusion, and numerator criteria and the evaluated results as determined from patient data. This listing is then transformed into summary statistics showing the number of patients who qualified for the measure and what proportion of these met the performance measure (i.e. met the numerator criterion). Secondary analyses showing various percentages of patients meeting certain exclusion criteria or failing to meet certain inclusion criteria are also computed to facilitate understanding of various contributing factors.

In contrast, the CDS output is a patient-specific advisory geared towards providers with the assumption that they will review the information and apply their clinical judgment. It is in a layered graphical user interface. The top level contains patient data, whether a patient reached the guideline goal, alerts, and detailed therapeutic recommendations. Therapeutic recommendations contain multiple guideline-compliant choices, each with patient specific indications, contraindications and adverse events highlighted. In a second layer we provide additional drug-related information such as the need for monitoring, drug dosing and other relevant alerts. The objective of the design is to bring the most important information for the decision making process upfront and to display additional relevant information upon user demand.

Analysis 2: Use Guideline-Based CDS to Complement Performance Measure Evaluation

To assist clinicians providing best practices in the performance measure environment we modified ATHENA CDS, originally developed as a standalone CDS system, into ATHENA PMtoCDS, a CDS system that provides advisories for patients failing performance measures. In the following we describe the changes made so that ATHENA PMtoCDS can play this role.

1. Workflow Integration

Instead of triggering CDS at the point of care, ATHENA PMtoCDS is designed to display pre-computed CDS recommendations when a provider reviews a panel of patients on the VISN dashboard. The software that computes dashboard performance measure is configured to generate a set of cases that fail the performance measures and to

pass the set to ATHENA PMtoCDS, which then generate the CDS. The dashboard displays patients' performance with respect to VA performance measures as a population-based summary and as individual patient records. For a patient who fails to meet selected performance measures, a user can bring up ATHENA PMtoCDS recommendations that suggest how to manage the treatment of that patient.

2. Cohort Definitions

Because the performance measure system filters out patients who satisfy performance measure targets, CDS recommendations will not be generated for those patients. As originally implemented, ATHENA CDS provides recommendations regarding all patients independent of their treatment goal. For example, in the ATHENA-Hypertension CDS system, if a patient's blood pressure is within the target range the system may recommend substitution from less desired to more preferred antihypertensive agents. These recommendations would never be generated in the performance measure environment since only patients with blood pressures above target would be eligible for the CDS. Another example is the way CDS manages glycemic control in patients with type II diabetes. The VA performance measure for the target HbA1c is less than 9% while in a point-of-care CDS system the target would set by the provider in conjunction with the patient and can change over time. The cohort of patients receiving ATHENA PMtoCDS recommendations is restricted to those whose HbA1c is greater than 9%.

3. Definition of Compliance

As described above, having ATHENA PMtoCDS work in conjunction with the performance measure dashboard means changing the HbA1c goal from an individualized, patient-provider agreed-upon target HbA1c to a single HbA1c goal (of less than 9%). Because recommendations and messages are given only to those who fail this performance measure, and are above this threshold, there are recommendations that would either not be given, or given only regarding a subset of patients rather than all patients. For example, if a patient has a HbA1c<6%, and meets certain other criteria the CDS would issue a message that he/she is potentially at increased risk for cardiovascular events; this message would not be issued when considering only those patients with a high HbA1c. Similarly, if a patient's bicarbonate level is less than or equal to 21 mEq/L, an alert would normally be given regarding any patient, regardless of their HbA1c level. Thus because of the pre-filtering of cohort receiving CDS, some guideline recommendations or alerts become inapplicable or are applied to a subset of patients.

4. Use of Data

The differences described in the results of Analysis 1 in how performance measures and CDS use patient data also apply to Analysis 2. Integration with the dashboard, however, enables the CDS to access data that are computed by the dashboard. This allowed for the implementation of CPG recommendations that could not be implemented before. For example, the dashboard computes the medication possession ratio (MPR) that indicates the level of adherence to the prescribed medications. If the MPR is less than 0.9, then ATHENA PMtoCDS issues a message, alerting the provider that the patient may not to be adhering to his/her therapy and re-evaluation of clinical strategy may be warranted. A recommendation to increase dose or to add addition medications may not be appropriate if a patient is not adhering to the prescribed regimen of existing medications.

5. Output Formats

Unlike the previous versions of ATHENA CDS, which were directly integrated with the VA's CPRS, the display is generated using the same Microsoft Report Server tool that generates the dashboard performance results. The advantage of using the same Report Server tool for display is that consistency of text font, size, colors, and navigation icons, i.e. the look and feel, between the CDS display and other dashboard pages can more easily be attained. Even then, to minimize possible confusion it is necessary to ensure that displayed terms from the ATHENA PMtoCDS output are made consistent with their counterparts in the dashboard.

The findings of the two analyses are summarized in Table 1. The 'Analysis 1' columns contrast the HF-PM and HF-CDS applications along the dimensions of workflow, cohort definition, definition of compliance, use of data, and output format. The 'Analysis 2' column indicates the changes to the CDS system so that it supplements a dashboard application that computes and displays performance measure information. For those patients who fail to satisfy specific performance measures, ATHENA PMtoCDS recommends changes to the management of the patients according to clinical practice guidelines relevant to the performance measures.

Table 1. Major findings: (1) Comparison of HF-PM and HF-CDS systems (Analysis 1) and (2) Necessary changes to ATHENA PMtoCDS for it to supplement a dashboard application that computes and displays performance measure information (Analysis 2).

	Analysis 1		Analysis 2
	HF-PM	**HF-CDS**	**Necessary Changes to PMtoCDS**
Workflow	Retrospective/on demand	Prospective/event driven	Triggered as part of PM Dashboard for patients failing PM
Cohort Definition	More restrictive to improve specificity; retrospective data more available	Less restrictive to allow clinical judgment; Data less available; Exclude patients who do not need recommendation; Exclude cases requiring complex management	Limited to patients failing PM
Definition of Compliance	Single numerator metric	More nuanced definition of best practice	Modified to be consistent with PM Dashboard; No recommendations for those who pass PMs
Use of Data	Stricter temporal limits to improve specificity	More inclusive	Able to use Dashboard analytics (medication possession ratio) to enhance CDS
Output Format	Population-based summary statistics	Patient-specific	Made to be consistent with that of PM Dashboard

Discussion

We use the implementations of two pairs of performance measure and CDS systems to demonstrate concretely how evaluating performance measures and CDS differ and yet can complement each other. In the first analysis, we compare the implementations of similar performance measures and guideline recommendations in the same computing environment, and systematically analyze how the resulting systems differ in workflow, definition of cohort, definition of compliance, use of data, and output formats. We attribute the differences in cohort definitions to discrepancies in the sources of the performance measures, the retrospective/prospective differences of performance measures and CDS, and, above all, in the role of clinical judgment in providing CDS. While performance measures need to be applied more narrowly to a denominator population for which the numerator criteria unambiguously apply, CDS recommendations can be provided to target patients whose management is entrusted to providers who can exercise their judgment based on data and knowledge that may not be available to the CDS system.

Analysis 2 explores the necessary changes to a CDS system for it to complement the performance measure services provided by the dashboard. We see how the pre-filtering of patient eligibility limits the recommendations that a CDS system can provide and how the consistency requirement between the performance measure system and CDS system forced the CDS system to re-define the therapeutic targets to achieve and to modify its output format.

In Analysis 1 we repurposed the ATHENA CDS infrastructure to compute performance measures. Even though CDS and performance measure program appear, on the surface, to be quite different because they produce different outputs, both programs are based on the recommendations from the same clinical evidence. Seen in this way, the underlying building blocks for both are either the same or very similar. Both make use the same medical conditions, medication lists, ICD9 codes, laboratory measurements and numeric cutoff criteria. Because we had previously encoded the heart-failure recommendations in a Protégé knowledge base, these building blocks, in the forms of Protégé classes, formalized criteria, and mappings from VA data sources, were already available for reuse when we encode performance measures for the heart-failure performance measure project. Having control of the execution engine also allows us to annotate the computed results with explanations (e.g., specific patient data causing a patient to fail an inclusion criterion) that help a user understand the computed performance measures. The disadvantage of repurposing the CDS infrastructure is that ATHENA CDS is designed to provide decision support for individual patients. It suffers from performance issues when computation is required for a very large patient cohort. Using SQL to implement performance measures, as the VISN 21 dashboard does, means that queries are applied to sets of patient data at a time, making the computation much more efficient.

Placing CDS in the service of improving the achievement of performance measures is not the only way to relate CDS and performance measures. In the literature we see additional possibilities. Fonarow et al.[18] used performance measures to measure the effectiveness of guideline-based CDS. They showed improvements in 5 out of

7 process performance measures among HF patients after 24 months of intervention that included clinical decision support tools, structured improvement strategies, and chart audits with feedback. LeBresh et al. [19] developed a "Get With The Guidelines" program that integrated a patient management tool with performance measures and guideline summaries. They found significant improvements from baseline to the fourth quarter in 11 of 13 measures[19]. Neither the Fonarow et al. nor LeBresh et al. studies used CDS that provides patient-specific management recommendations generated from automated CPG. Walter at al [20] used the best practices defined in CPGs to create performance measure. Using colorectal cancer screening as a case study, they identified a number of pitfalls in using CPGs to define performance measures that are similar to our findings about the differences between CDS and performance measures, such as not accounting for provider judgment when scoring performance measures. Finally, van Gendt et al. [21] used performance measures developed independently of CPGs to evaluate completeness of CPGs and to offer the possibility of improving CPGs. The authors formalized performance measures as goals that CPG recommendations should satisfy. For example, a performance measure may measure the percentage of diabetes patients with albumin value measured in a 12-month period. They analyzed a Type-2 DM CPG to see whether the paths in the CPG would achieve the goal of having an albumin measurement within that period. They found that, out of 35 performance measures studied, 25 (71%) suggested that there are problems with CPGs they used in the study.

Conclusion

Our analyses demonstrate that (1) the same evidence-based treatment recommendations are implemented differently as performance measures and as management advice in a CDS system and (2) a CDS system can help patients achieving compliance with performance measures but would require significant modifications so that it complements a performance dashboard in a consistent manner.

Acknowledgement

The work was supported by VA HSR&D grant IIR 11-071 and VA HSR&D Quality Enhancement Research Initiative (QUERI) grants RRP 11-428 and RRP 12-447. Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs or other affiliated institutions.

References

1. Jha AK, Perlin JB, Kizer KW, Dudley RA. Effect of the transformation of the Veterans Affairs Health Care System on the quality of care. N Engl J Med. 2003 May 29;348(22):2218-27. PubMed PMID: 12773650.
2. Kizer KW. The "new VA": a national laboratory for health care quality management. Am J Med Qual. 1999 Jan-Feb;14(1):3-20. PubMed PMID: 10446659.
3. U.S. Department of Veteran Affairs. VA/DoD Clinical Practice Guidelines [3/1/2016]. Available from: http://www.healthquality.va.gov.
4. Goldstein MK, Hoffman BB, Coleman RW, Musen MA, Tu SW, Advani A, et al. Implementing clinical practice guidelines while taking account of changing evidence: ATHENA DSS, an easily modifiable decision-support system for managing hypertension in primary care. Proc AMIA Symp. 2000:300-4. PubMed PMID: 11079893. Pubmed Central PMCID: 2243943.
5. Goldstein MK, Hoffman BB, Coleman RW, Tu SW, Shankar R, O'Connor M, et al. Offline Testing of Automated Decision Support for Management of Hypertension HSR&D Annual Meeting. 2001.
6. Goldstein MK, Hoffman BB, Coleman RW, Tu SW, Shankar RD, O'Connor M, et al. Patient safety in guideline-based decision support for hypertension management: ATHENA DSS. Proc AMIA Symp. 2001:214-8. PubMed PMID: 11825183. Pubmed Central PMCID: 2243380. Epub 2002/02/05. eng.
7. Goldstein MK, Coleman RW, Tu SW, Shankar RD, O'Connor MJ, Musen MA, et al. Translating research into practice: organizational issues in implementing automated decision support for hypertension in three medical centers. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):368-76. PubMed PMID: 15187064. Pubmed Central PMCID: 516243.
8. Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubezy M, Eriksson H, et al. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. Int J Hum Comput Stud. 2003;58(1):89-123.

9. Michel M, Trafton J, Martins S, Wang D, Tu S, Johnson N, et al. Improving Patient Safety Using ATHENA-Decision Support System Technology: The Opioid Therapy for Chronic Pain Experience. In: Henriksen K, Battles JB, Keyes MA, Grady ML, editors. Advances in Patient Safety: New Directions and Alternative Approaches (Vol 4: Technology and Medication Safety). Advances in Patient Safety. Rockville (MD)2008.
10. Trafton J, Martins S, Michel M, Lewis E, Wang D, Combs A, et al. Evaluation of the acceptability and usability of a decision support system to encourage safe and effective use of opioid therapy for chronic, noncancer pain by primary care providers. Pain Med. 2010 Apr;11(4):575-85. PubMed PMID: 20202142. Epub 2010/03/06. eng.
11. Trafton JA, Martins SB, Michel MC, Wang D, Tu SW, Clark DJ, et al. Designing an automated clinical decision support system to match clinical practice guidelines for opioid therapy for chronic pain. Implement Sci. 2010;5:26. PubMed PMID: 20385018. Pubmed Central PMCID: 2868045.
12. National Quality Forum. NQF-Endorsed Measures (QPS) 2012 [last accessed 2014 March 13]. Available from: http://www.qualityforum.org/Measures_Reports_Tools.aspx.
13. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, et al. 2013 ACCF/AHA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation. 2013 October 15, 2013;128(16):e240-e327.
14. Tu SW, Musen MA. Modeling data and knowledge in the EON guideline architecture. Stud Health Technol Inform. 2001;84(Pt 1):280-4. PubMed PMID: 11604749. Epub 2001/10/18. eng.
15. Tu SW, Musen MA, editors. A Flexible Approach to Guideline Modeling. Proc AMIA Symp; 1999; Washington DC: Hanley & Belfus, Inc.
16. Shiffman RN, Michel G, Essaihi A, Thornquist E. Bridging the Guideline Implementation Gap: A Systematic, Document-Centered Approach to Guideline Implementation. J Am Med Inform Assoc. 2004;11:418-26.
17. Tso GJ, Tu SW, Oshiro C, Martins S, Wang D, Robinson A, et al. Automating Guidelines for Clinical Decision Support: Knowledge Engineering and Implementation. AMIA Symposium; Chicago 2016. Submitted.
18. Fonarow GC, Albert NM, Curtis AB, Stough WG, Gheorghiade M, Heywood JT, et al. Improving evidence-based care for heart failure in outpatient cardiology practices: primary results of the Registry to Improve the Use of Evidence-Based Heart Failure Therapies in the Outpatient Setting (IMPROVE HF). Circulation. 2010 Aug 10;122(6):585-96. PubMed PMID: 20660805.
19. LaBresh KA, Gliklich R, Liljestrand J, Peto R, Ellrodt AG. Using "get with the guidelines" to improve cardiovascular secondary prevention. Jt Comm J Qual Saf. 2003 Oct;29(10):539-50. PubMed PMID: 14567263.
20. Walter LC, Davidowitz NP, Heineken PA, Covinsky KE. Pitfalls of converting practice guidelines into quality measures: lessons learned from a VA performance measure. JAMA. 2004 May 26;291(20):2466-70. PubMed PMID: 15161897.
21. van Gendt M, ten Teije A, Serban R, van Harmelen F. Formalising medical quality indicators to improve guidelines. AI in Medicine, 2005.