NIH Public Access
Author Manuscript
Med Care. Author manuscript; available in PMC 2014 March 01.

Published in final edited form as:
Med Care. 2013 March ; 51(3): 266–274. doi:10.1097/MLR.0b013e31827da99c.

Improving the Reliability of Physician "Report Cards"

Kimberly A. Smith, MD, MS[1,2,3], **Jeremy B. Sussman, MD, MS**[2,4], **Steven J. Bernstein, MD, MPH**[2,4], and **Rodney A. Hayward, MD**[1,2,4]

Kimberly A. Smith: kcandido@umich.edu; Jeremy B. Sussman: jeremysu@umich.edu; Steven J. Bernstein: sbernste@umich.edu

[1]Robert Wood Johnson Foundation Clinical Scholars Program, VA Health Services Research & Development Center of Excellence, VA Ann Arbor Healthcare System, Ann Arbor, Michigan

[2]University of Michigan Divisions of General Internal Medicine, VA Health Services Research & Development Center of Excellence, VA Ann Arbor Healthcare System, Ann Arbor, Michigan

[3]Nephrology, VA Health Services Research & Development Center of Excellence, VA Ann Arbor Healthcare System, Ann Arbor, Michigan

[4]Center for Clinical Management Research, VA Health Services Research & Development Center of Excellence, VA Ann Arbor Healthcare System, Ann Arbor, Michigan

Abstract

Background—Performance measures are widely used to profile primary care physicians (PCPs) but their reliability is often limited by small sample sizes. We evaluated there liability of individual PCP profiles and whether they can be improved by combining measures into composites or by profiling practice groups.

Methods—We performed a cross-sectional analysis of electronic health record data for patients with diabetes (DM), congestive heart failure (CHF), ischemic vascular disease (IVD), or eligible for preventive care services seen by a PCP within a large, integrated healthcare system between April 2009 and May 2010. We evaluated performance on 14 measures of DM care, 9 of CHF, 7 of IVD, and 4 of preventive care.

Results—There were 51,771 patients seen by 163 physicians in 17 clinics. Few PCPs (0 to 60%) could be profiled with 80% reliability using single process or intermediate-outcome measures. Combining measures into single-disease composites improved reliability for DM and preventive care with 74.5% and 76.7% of PCPs having sufficient panel sizes, but composites remained unreliable for CHF and IVD. 85.3% of PCPs could be reliably profiled using a single overall composite. Aggregating PCPs into practice groups (3 to 21 PCPs per group) did not improve reliability in most cases due to little between-group practice variation.

Conclusion—Single measures rarely differentiate between individual PCPs or groups of PCPs reliably. Combining measures into single- or multi-disease composites can improve reliability for

Rodney A. Hayward, MD (corresponding author and requests for reprints), 6312 Medical Science Building I University of Michigan Health System Ann Arbor, MI 48109-5604 (734) 647- 3301 rhayward@umich.edu.
Complete Author Information, Kimberly A. Smith, MD, MS, 7075 Garden Walk, Columbia, MD 21044 (410) 531-3535, **Jeremy B. Sussman, MD, MS**, 2800 Plymouth Rd Bldg 16/441E North Campus Research Complex University of Michigan Health System Ann Arbor MI 48109-2800 (734) 936-5291, **Steven J. Bernstein, MD, MPH**, 2800 Plymouth Rd Bldg 16/446E North Campus Research Complex University of Michigan Health System Ann Arbor MI 48109-2800 (734) 936-5213

Supplemental Digital Content: Supplemental Digital Content 1.docx

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

some common conditions, but not all. Assessing PCP practice groups with in a single healthcare system, rather than individual PCPs, did not substantially improve reliability.

Keywords

quality measurement; reliability; physician profiling

Introduction

Amidst growing concerns about the quality of American healthcare, many health systems now measure physician performance to inform quality improvement efforts and to increase accountability for achieving quality goals.(1-3) Such initiatives often target primary care providers' (PCPs) care of patients with chronic illness, as this care has the potential to impact a growing proportion of morbidity and cost.(1, 4, 5) These initiatives often use performance measures, such as annual diabetic foot examinations or antiplatelet agent use in patients with known ischemic vascular disease. Each individual measure, which assesses a specific type of care for a single disease, produces a result that reflects a very narrow aspect of a provider's practice. While consumers sometimes value these granular data, composite measures may provide a broader perspective on quality. In addition, for optimal provider profiling, measures should be able to reliably detect meaningful variation between the individuals being compared. If the number of patients is in adequate or the variability between physicians is too low, performance profiles may mislead. The conventional expectation is that quality profiles should achieve at least 80% reliability when judging individuals (such as “grading” hospitals or doctors), and even higher levels are generally recommended when judgments have serious consequences (such as strong financial penalties).(6) Unfortunately, several studies suggest that existing profiling methods often lack this critically important level of reliability.(7-11)

The reliability of a quality profile could be improved with a larger sample size, such as by adding more measures or more patients, or by profiling an entire practice group rather than an individual physician. Reliability can also be improved by better identifying areas of between-physician variation, such as with more accurate quality measures or more reliable data sources. In this study, we use electronic health record (EHR) data to evaluate two possible ways to improve the reliability of performance profiles in a large, integrated healthcare system: combining measures into single-disease or multi-disease composites and combining PCPs into practice groups to increase sample size.(12-15)

Methods

Setting and Participants

We conducted this study within an academic-affiliated healthcare system comprised of 3 hospitals, 40 ambulatory care sites, and more than 1,600 physicians providing over 1.7 million clinic visits per year. There are 163 PCPs and geriatricians, most of whom are full-time clinicians practicing in community settings.

We identified all patients between ages 18 and 75 who were seen in a family medicine, general medicine, medicine/pediatrics, or geriatrics outpatient clinic between April 2009 and May 2010. We attributed patients to a specific PCP if they had been seen by the provider at least twice in the preceding 2 years, including once in the past 13 months. While geriatricians can function as consultants or PCPs, for the purposes of institutional disease registries and feedback reports, patients are assigned to geriatricians as PCPs if they meet the above attribution criteria.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

We then limited the population to patients eligible for preventive care services and patients included in registries for diabetes (DM), congestive heart failure (CHF), and ischemic vascular disease (IVD.) The registries are designed to measure core quality indicators for physician feedback, internal quality improvement, and reporting to external programs.

Patients with DM were identified by 1 of 3 criteria: (1) 2 diagnoses of diabetes in billing data in an ambulatory care setting in the past 3 years; (2) 1 diagnosis of diabetes in billing data in an acute care setting such as the emergency department or inpatient setting in the past 3 years; or (3) a diagnosis of diabetes in the EHR problem summary list (PSL). Validation of the diagnosis required 1 of the following criteria: (1) diabetes documented in a clinical note; (2) a prescription for hypoglycemic medication (excluding metformin alone); (3) a prescription for diabetic supplies; (4) hemoglobin A1c> 6.4%; or (5) 2 or more blood glucose levels ≥ 200 mg/dL.

Patients with CHF were identified by one of the following 4 criteria: (1) 2 diagnoses of heart failure in billing data in an ambulatory care setting; (2) 1 diagnosis of heart failure in an acute care setting; (3) a diagnosis of heart failure in the EHR PSL; or (4) 1 diagnosis of heart failure in an ambulatory care setting and either evidence of low ejection fraction or heart failure documented in EHR PSL. Criteria 1, 2, and 3 required validation by one of the following: admission in the past 2 years with a principal diagnosis of heart failure, a Brain Natriuretic Peptide (BNP) ≥ 100 in the past 2 years, an ejection fraction (EF) < 40% on echocardiogram or nuclear medicine test, heart failure recorded on the EHR PSL, or a reference to New York Heart Association classification in the EHR within the preceding 2 years. Patients were excluded if they had a history of heart transplant, left ventricular assist device, or congenital heart disease.

Patients with IVD were identified by 1 of the following criteria: inpatient encounter for coronary artery bypass grafting (CABG), percutaneous transluminal coronary angioplasty (PTCA), percutaneous transluminal coronary intervention (PCI), stroke, transient ischemic attack (diagnosed on the in-patient neurology service only), acute myocardial infarction (AMI) or an EHR PSL entry for CABG, PTCA, PCI or AMI. Patients were excluded if they had a history of heart transplant or pulmonary hypertension.

This study was approved by the University of Michigan Institutional Review Board.

Outcomes and Follow-up

We obtained detailed patient-level data including demographics from the health system's EHR and billing records. This included laboratory and immunization data, mammography findings, physical exam findings such as blood pressure and the date of the most recent eye and foot examinations and ejection fraction. We obtained medications including contraindications from the EHR medication summary list.

We identified performance measures from an environmental scan of existing publicly-available physician measures (e.g. Healthcare Effectiveness Data and Information Set (HEDIS,)(16) Rand Health Quality of Care Assessment Tool,(17) American College of Cardiology/American Heart Association measures.(18, 19)) When possible, we based measures on those currently used within the health system to provide feedback to providers and to inform internal quality improvement initiatives. The measure topics can be found in Table 1 (see Table, Supplemental Digital Content 1, for a more detailed description of the measure criteria.) We converted all measures into dichotomous (adherent/not adherent) outcomes for analysis.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Statistical Analysis

We analyzed the data using multilevel logistic regression techniques in Stata version 11.1. This method accounts for the clustered nature of the data, i.e., multiple measures for an individual patient, multiple patients cared for by a single physician, and several physicians practicing together at the same site (PCP group). First, we developed separate models for each measure. We started with unadjusted random intercept models, i.e., “empty” models, to partition variance in performance measures between the different levels of the hierarchy. Next, we added fixed-effect patient-level variables for case-mix adjustment, including patient age, sex, and number of eligible chronic illnesses (0 to 3). Variance estimates from these models allowed us to calculate the intra class correlation coefficient (ICC). The ICC represents the fraction of total variability attributable to a particular level of the model (patient, PCP, or PCP group). Reliability is the ability of a measure to provide the same result on repeat testing. The ICC at the PCP level accounts for the correlation in patient measurements within an individual PCP as an estimate of the PCP effect on those measurements. The ICC at the PCP group level accounts for the correlation in patient measurements within an individual PCP, and for PCP measurements within a PCP group as an estimate of the PCP group effect on those measurements.

We then developed single-disease composite models using all measures related to a specific disease that had a non-zerocase-mix-adjusted PCP-level ICC in the single-measure models. Since physicians vary in the types of patients they see and specific measures are associated with higher or lower physician performance, we included a fixed-effect measure-level variable: the proportion of eligible patients for whom the measure was achieved in the overall population. Finally, we developed an overall multi-disease composite model using all measures included in the single-disease composite models. For all composite measures, we equally weighted each individual measure observation as has been described previously for measures assumed to have equal validity and importance.(20)

We used posterior empirical Bayes estimates from the single- and multi-disease composite models to predict PCP quality scores that reflect their expected performance on an average-difficulty measure for a 65 year-old female patient with 1 comorbidity. This produced case-mix-adjusted PCP profiles that were also “reliability-adjusted” (often referred to as “shrunken” estimates) to account for bias resulting from suboptimal reliability.(7, 21) To compare differences between physician profiles, we report results with confidence intervals of 1.4 standard errors (~83% confidence intervals). This is the level at which overlap of two estimates' confidence intervals indicates that the differences between the results are not statistically significant at the α=0.05 level.(22, 23)

We applied the Spearman-Brown prophecy formula to determine the panel size needed to achieve a measurement reliability of 80%. This suggests that 80% of the variation in a profile is due to differences in practice and 20% is due to chance, which is often considered the minimum reliability needed to make decisions about individuals.(6, 7) We reported the percentage of providers who had a panel size that would fulfill this criterion.(6-8, 11)

Results

The final study sample included 51,771 patients cared for in 17 clinics by 1 of 163 primary care providers (PCPs). PCPs had a mean number of 51 (range 1-219) patients eligible for the DM registry, 8 (1-34) for CHF, 19 (1-107) for IVD, and 220 (1-837) for preventive care. The number of PCPs per clinic ranged from 3 to 21. Table 1 summarizes the overall adherence with each performance measure across the population. The proportion of patients meeting a measure ranged from a low of 45.5% of patients with diabetes attaining a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

hemoglobin A1c of <7% to a high of 98.5% of CHF patients having ejection fraction (EF) assessed.

Variation between PCPs was statistically significant for all single measures of prevention and most of DM (Table 2). In contrast, there was no significant variation in performance between PCPs for any individual CHF or IVD measure. For prevention, approximately 60% of PCPs had sufficient patient panel sizes to be profiled with 80% reliability for both breast and cervical cancer screening whereas less than 20% could be reliably profiled using the pneumococcal and influenza immunization measures. Few PCPs (0 to 40%) had sufficient panel sizes to be profiled with 80% reliability on the basis of any individual DM process measure and none on the basis of any DM intermediate-outcome measure (i.e., hemoglobin A1c level or blood pressure control).

After combining measures into single-disease composites, variation between PCPs was statistically significant for all but the CHF composite. For diabetes and preventive care, 74.5% and 76.7% of PCPs respectively had sufficient panel sizes to be profiled with 80% reliability (Table 3). However, a mere 2.1% of PCPs had sufficient panel sizes for IVD care. After combining measures into an overall multi-disease composite, 85.3% of PCPs had sufficient patients to be profiled with 80% reliability (Figure 1).

Variation between PCP groups (practice sites) was sufficient to reliably distinguish their performance on measures of breast and cervical cancer screening. However, there was no significant variation in performance between PCP groups on single measures for DM, CHF, IVD, or immunizations (Table 2). All PCP groups had sufficient patient panel sizes to be profiled with 80% reliability for breast cancer screening and 94% for cervical cancer screening.

After combining measures into single-disease and an overall composite, 100% of PCP groups could be profiled with 80% reliability for the preventive care and overall composites, although the ICCs for PCP groups were consistently much lower than those for individual PCPs. In contrast, there was no significant variation between PCP groups for composites of DM, CHF, or IVD care (Table 4).

Discussion

Our study provides valuable information on proposed strategies for improving the reliability of PCP profiling within a large, integrated healthcare system. We found that few PCPs or PCP groups could be reliably profiled using single process or intermediate-outcome measures of diabetes, congestive heart failure, ischemic vascular disease, and preventive care. When we combined individual measures into disease-based composites, reliability improved substantially, primarily due to the increased number of measures and/or eligible patients for the more common conditions of DM and preventive care. Composites remained unreliable for CHF and IVD due to a combination of them being less common, having fewer measures and there being less variation between PCP scores. Although profiling PCP practice groups instead of individual PCPs can greatly increase sample size, with the exception of preventive care profiling, PCP groups generally lowered reliability because there was more variation between individual PCPs than there was between PCP practice groups.

Previous studies on the reliability of individual performance measures have yielded mixed results.(11) In some instances, measures of patient satisfaction, preventive care, or chronic care have demonstrated sufficient variability between physicians or groups for reliable profiling.(8, 24-26) More commonly, individual performance measures have proven unreliable.(5-10, 26-31). These disparate findings are not necessarily contradictory but

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

rather demonstrate that the degree and sources of performance variation can vary by contextual factors such as practice type, setting, or even over time. Appropriate consideration of context is known to be critical for the successful spread of quality improvement interventions.(32, 33) Similarly, contextual factors are likely to influence the usefulness of performance profiles, highlighting the need to use methods, such as those used in this paper, to evaluate the reliability of measures for a specific use prior to linking profiles to rewards or penalties. A profiling system that has proven reliable in one setting may prove unreliable in another.

In some instances, combining measures into composites has improved reliability.(14, 29) In our study, single-disease composites dramatically increased our ability to distinguish between PCPs for both DM and preventive care, with approximately 75% of PCPs having enough patients with these conditions for reliable profiling. For PCP groups, 100% could be reliably profiled with the preventive care composite. In contrast, single-disease composites did not improve reliability for either CHF or IVD. Relative to DM and preventive care, physicians had relatively few patients with CHF or IVD, there were fewer measures for these diseases, and there was little variation between providers in achieving these measures (i.e., “topped-out” measures with high overall performance). This high performance may have been the result of intense measurement and quality improvement activities in these areas prior to our study. Patients with CHF or IVD may also be more likely to be co-managed by sub-specialty consultants, making the PCP or PCP group a less relevant focus of profiling efforts, although we were not able to evaluate this hypothesis directly. Regardless, given the burden of illness and attention received by major conditions like CHF and IVD, performance profiles are often used to drive care improvement. Our data provide a cautionary note, however, that the value of these profiles to differentiate between PCPs or practice groups may be limited by poor reliability.

Aggregating measures one step further by combining all measures included in the single-disease composites into an overall composite increased reliability at the PCP level, by increasing sample size *and* increasing the variation in scores between PCPs. Although the composite score also allowed reliable profiling of PCP practice groups, this was due to increased sample size, not due to greater variation between PCP practice groups. Further, the overall composite improved reliability mainly due to combining the two most reliable condition-specific measures, DM and preventive care. This improved reliability could be negated if performance on one measure or one disease poorly correlates with performance in other areas.(34) In this situation, a signal of poor performance could be masked by high performance in other areas. Despite these potential limitations of composites, they did improve the reliability of our quality profiles and may be useful for high-level comparisons between providers or groups. In contrast, the results from single measures may provide the granularity necessary for targeted quality improvement activities in situations where differentiating between physicians is not the primary intent and limitations in measure reliability can be taken into consideration.

In general, for individual measures, single-disease composites, and the overall composite, aggregating providers into practice groups did not improve reliability, despite the large increase in sample size, because between-group variation was insignificant for most measures. Since the provider groups practice at different sites within a single integrated healthcare system, they can be considered clinical micro systems.(35, 36) It is possible that broader institutional efforts have reduced variability between micro systems through access to similar resources (e.g., the same EHR), administrative structure, quality improvement initiatives, and comparative measurement targeted to the group level. While this may have resulted in greater consistency between practice sites, variation persists between individual providers that are not specifically targeted. Others studying more diverse populations have

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

reported greater variability between practice groups, (8, 24-26) once again demonstrating the need to consider context and goals of measurement, and to determine prospectively whether a profiling system can meet those goals.

A significant strength of our study was the ability to use an electronic health record system, allowing us to pool data from patients regardless of health insurance and to obtain more detailed data than administrative datasets allow, including laboratory results, medication allergies, and other measure exclusions. Electronic health record data also improve attribution of patients to specific physicians, aid assignment of disease categories, and facilitate the assessment of quality measures.(37)

There are several limitations to our study. Although most of the physicians practice full-time in community settings, they were associated with a single academic healthcare system and were exposed to similar quality improvement activities including feedback reports, educational initiatives, and incentives to improve on the measures used in this study. These activities may have improved measure performance and reduced variation. Our results may also not be generalizable to a system with a less robust medical informatics infrastructure or to measures based on less detailed clinical data, such as those lacking exclusions for contraindications. Given the national push to extend both this infrastructure and quality measurement, however, our results may better inform future efforts to use electronic health records to assess and improve quality. As in many performance measurement studies, attributing patients to a specific PCP was often difficult and undoubtedly not always accurate. Also, patients who received care elsewhere during the study may not have had all results captured in the EHR. Although missing data could decrease measure performance in general, its impact on variability is uncertain. Most health systems and performance measurement efforts face similar challenges. Finally, our results relate to reliability and not necessarily validity. Validity requires that what is being measured truly represents differences in quality. Although we used standard measures in widespread use, there remains a need to further improve the clinical accuracy and salience of performance measurement as well as the ability to account for patient preferences or characteristics that could drive differential selection of providers or groups.(38, 39) Lastly, while our study cannot fully explain why aggregating measurement to the PCP group level or by creating composites does not improve reliability for each condition, we offer several hypotheses that may be tested in future analyses.

Performance profiles are now a common component of programs to encourage informed patient decision-making and drive accountability for quality improvement. It is critical that these profiles provide valid and reliable assessments of quality. Our study demonstrates that single measures will frequently produce unreliable profiles for both individual PCPs and for PCP group practices with in a single healthcare system. Combining these measures into single-disease or multi-disease composites may improve reliability for some common diseases (such as diabetes) or types of care (such as preventive care), but not for other important conditions (such as CHF and IVD).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding sources: This work was supported in part by the Robert Wood Johnson Foundation Clinical Scholars Program, the Department of Veteran Affairs Quality Enhancement Research Initiative (QUERI), and the Measurement Core of the Michigan Diabetes Research & Training Center (NIDDK of The National Institutes of Health [P60 DK-20572]).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

References

1. Christianson JB, Volmar KM, Alexander J, et al. A report card on provider report cards: current status of the health care transparency movement. J Gen Intern Med. 2010; 25:1235–1241. [PubMed: 20625849]
2. Rosenthal MB, Landon BE, Normand SL, et al. Pay for performance in commercial HMOs. N Engl J Med. 2006; 355:1895–1902. [PubMed: 17079763]
3. The Commonwealth Fund Commission on a High Performance Health System. Health System Performance. 2011. Why Not the Best? Results from the National Scorecard on U.S.
4. Boyd CM, Darer J, Boult C, et al. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. JAMA. 2005; 294:716–724. [PubMed: 16091574]
5. Keating NL, Landrum MB, Landon BE, et al. The influence of physicians' practice management strategies and financial arrangements on quality of care among patients with diabetes. Med Care. 2004; 42:829–839. [PubMed: 15319608]
6. McDowell, I.; Newell, C. Measuring Health: A Guide to Rating Scales and Questionnaires. New York, NY: Oxford University Press; 2006.
7. Hofer TP, Hayward RA, Greenfield S, et al. The unreliability of individual physician “report cards” for assessing the costs and quality of care of a chronic disease. JAMA. 1999; 281:2098–2105. [PubMed: 10367820]
8. Krein SL, Hofer TP, Kerr EA, et al. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. Health Serv Res. 2002; 37:1159–1180. [PubMed: 12479491]
9. Huang IC, Diette GB, Dominici F, et al. Variations of physician group profiling indicators for asthma care. Am J Manag Care. 2005; 11:38–44. [PubMed: 15697099]
10. Turenne MN, Hirth RA, Pan Q, et al. Using knowledge of multiple levels of variation in care to target performance incentives to providers. Med Care. 2008; 46:120–126. [PubMed: 18219239]
11. Fung V, Schmittdiel JA, Fireman B, et al. Meaningful variation in performance: a systematic literature review. Med Care. 2010; 48:140–148. [PubMed: 20057334]
12. Mehrotra A, Adams JL, Thomas JW, et al. Cost profiles: should the focus be on individual physicians or physician groups? Health Aff (Millwood). 2010; 29:1532–1538. [PubMed: 20679658]
13. O'Connor PJ, Rush WA, Davidson G, et al. Variation in quality of diabetes care at the levels of patient, physician, and clinic. Prev Chronic Dis. 2008; 5:A15. [PubMed: 18082004]
14. Kaplan SH, Griffith JL, Price LL, et al. Improving the reliability of physician performance assessment: identifying the “physician effect” on quality and creating composite measures. Med Care. 2009; 47:378–387. [PubMed: 19279511]
15. Schmittdiel J, Vijan S, Fireman B, et al. Predicted quality-adjusted life years as a composite measure of the clinical value of diabetes risk factor control. Med Care. 2007; 45:315–321. [PubMed: 17496715]
16. National Committee for Quality Assurance. HEDIS 2010. Technical Specifications. 2009:2.
17. Asch SM, Kerr EA, Keesey J, et al. Who is at greatest risk for receiving poor-quality health care? N Engl J Med. 2006; 354:1147–1156. [PubMed: 16540615]
18. Drozda J Jr, Messer JV, Spertus J, et al. ACCF/AHA/AMA-PCPI 2011 Performance Measures for Adults With Coronary Artery Disease and Hypertension: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures and the American Medical Association-Physician Consortium for Performance Improvement. Circulation. 2011; 124:248–270. [PubMed: 21670226]
19. Bonow RO, Bennett S, Casey DE Jr, et al. ACC/AHA clinical performance measures for adults with chronic heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures (Writing Committee to Develop Heart Failure Clinical Performance Measures) endorsed by the Heart Failure Society of America. J Am Coll Cardiol. 2005; 46:1144–1178. [PubMed: 16168305]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

20. Asch SM, McGlynn EA, Hogan MM, et al. Comparison of quality of care for patients in the Veterans Health Administration and patients in a national sample. Annals of internal medicine. 2004; 141:938–945. [PubMed: 15611491]

21. Hayward RA, Heisler M, Adams J, et al. Overestimating outcome rates: statistical estimation when reliability is suboptimal. Health Serv Res. 2007; 42:1718–1738. [PubMed: 17610445]

22. Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: what do they mean in terms of statistical significance? J Insect Sci. 2003; 3:34. [PubMed: 15841249]

23. Goldstein H, Healy M. The Graphical Presentation of a Collection of Means. Journal of the Royal Statistical Society Series A (Statistics in Society). 1995; 158:175–177.

24. Safran DG, Karp M, Coltin K, et al. Measuring patients' experiences with individual primary care physicians. Results of a statewide demonstration project. J Gen Intern Med. 2006; 21:13–21. [PubMed: 16423118]

25. Solomon LS, Zaslavsky AM, Landon BE, et al. Variation in patient-reported quality among health care organizations. Health Care Financ Rev. 2002; 23:85–100. [PubMed: 12500472]

26. Sequist TD, Schneider EC, Li A, et al. Reliability of medical group and physician performance measurement in the primary care setting. Med Care. 2011; 49:126–131. [PubMed: 20421826]

27. Katon W, Rutter CM, Lin E, et al. Are there detectable differences in quality of care or outcome of depression across primary care providers? Med Care. 2000; 38:552–561. [PubMed: 10843308]

28. Orav EJ, Wright EA, Palmer RH, et al. Issues of variability and bias affecting multisite measurement of quality of care. Med Care. 1996; 34:SS87–101. [PubMed: 8792792]

29. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. Am J Manag Care. 2008; 14:833–838. [PubMed: 19067500]

30. Nyweide DJ, Weeks WB, Gottlieb DJ, et al. Relationship of primary care physicians' patient caseload with measurement of quality and cost performance. JAMA. 2009; 302:2444–2450. [PubMed: 19996399]

31. Rodriguez HP, Perry L, Conrad DA, et al. The Reliability of Medical Group Performance Measurement in a Single Insurer's Pay for Performance Program. Med Care. 2012; 50:117–123. [PubMed: 21993058]

32. Stevens DP, Shojania KG. Tell me about the context, and more. BMJ quality & safety. 2011; 20:557–559.

33. Dy SM, Taylor SL, Carr LH, et al. A framework for classifying patient safety practices: results from an expert consensus process. BMJ quality & safety. 2011; 20:618–624.

34. Palmer RH, Wright EA, Orav EJ, et al. Consistency in performance among primary care practitioners. Med Care. 1996; 34:SS52–66. [PubMed: 8792789]

35. Mohr JJ, Batalden PB. Improving safety on the front lines: the role of clinical microsystems. Quality & safety in health care. 2002; 11:45–50. [PubMed: 12078369]

36. Nelson EC, Batalden PB, Mohr JJ, et al. Building a quality future. Frontiers of health services management. 1998; 15:3–32. [PubMed: 10182606]

37. Kerr EA, Smith DM, Hogan MM, et al. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. The Joint Commission journal on quality improvement. 2002; 28:555–565. [PubMed: 12369158]

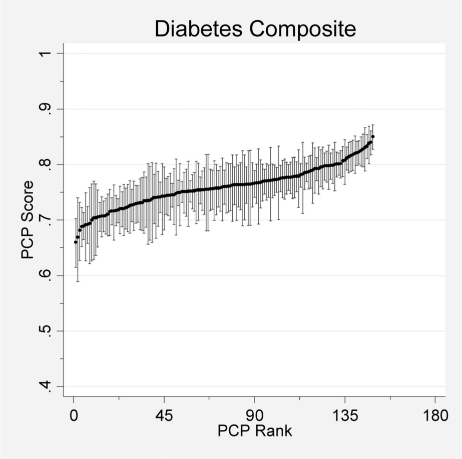
38. Hayward RA. Access to clinically-detailed patient information: a fundamental element for improving the efficiency and quality of healthcare. Med Care. 2008; 46:229–231. [PubMed: 18388837]

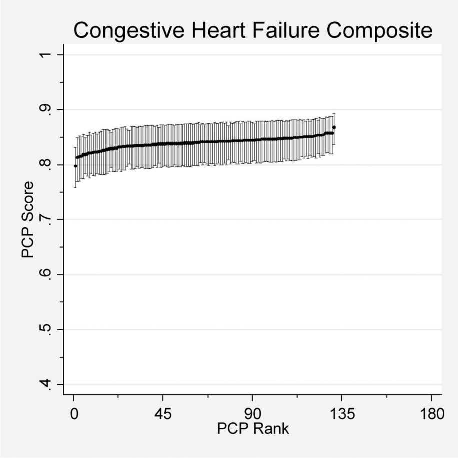
39. Kerr EA, Smith DM, Hogan MM, et al. Building a better quality measure: are some patients with ‘poor quality’ actually getting good care? Med Care. 2003; 41:1173–1182. [PubMed: 14515113]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

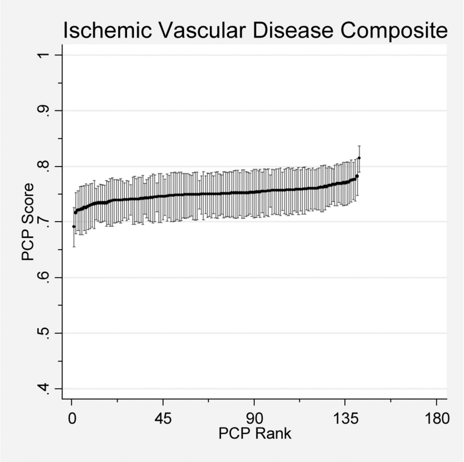


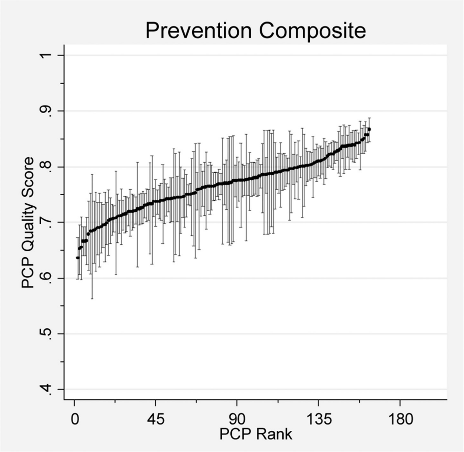


NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript





NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

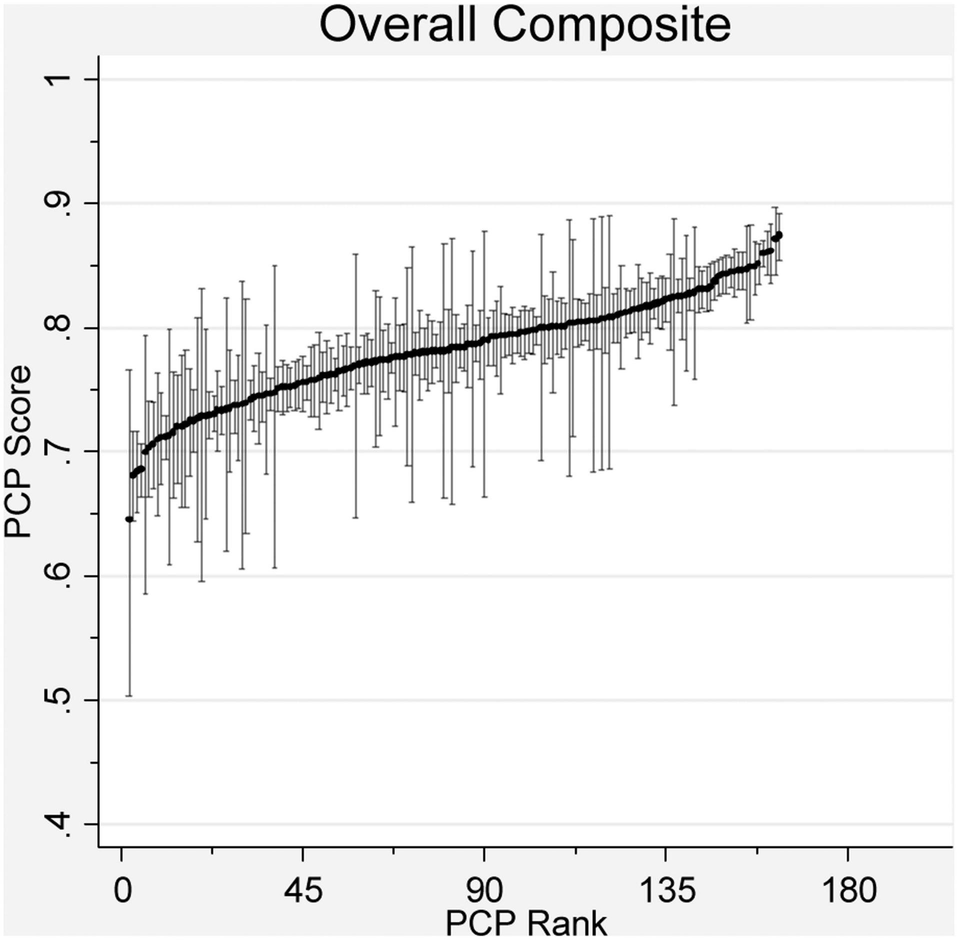


Figure 1. PCP Quality Scores

Reliability-adjusted primary care physician (PCP) quality scores with confidence intervals for single-disease and overall composite measures. Score reflects the expected PCP performance on an average difficulty measure for a 65 year-old female patient with one comorbidity.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Performance measure adherence.

Measures*	Type of Measure	#of Eligible Patients	Percent Meeting Measure
Diabetes			
ACE or ARB if proteinuric	Process	5,097	94.8%
Statin if indicated	Process	6,356	88.4%
Annual LDL	Process	7,628	85.7%
Semi-annual PCP evaluation	Process	7,628	84.9%
Semi-annual hemoglobin A1c	Process	7,628	82.6%
Annual foot examination	Process	7,628	70.6%
Pneumococcal vaccine	Process	7,628	68.9%
Influenza vaccine	Process	7,628	63.5%
Monitor for nephropathy	Process	2,393	61.8%
Annual eye examination	Process	7,628	60.8%
Antiplatelet agent if indicated	Process	6,807	59.6%
Hemoglobin A1c ≤ 7%	Intermediate Outcome	4,033	45.5%
Hemoglobin A1c ≤ 9%	Intermediate Outcome	3,195	92.2%
Blood pressure ≤ 135/80	Intermediate Outcome	7,628	63.9%
Congestive Heart Failure			
Ejection fraction (EF) assessed	Process	1,037	98.5%
Annual creatinine and potassium if on ACE or ARB	Process	717	93.0%
Statin if indicated	Process	823	87.5%
ACE or ARB if EF< 40%	Process	316	85.8%
Beta-blocker if EF< 40%	Process	336	83.6%
Annual LDL	Process	1,037	80.1%
Pneumococcal vaccine	Process	1,037	78.7%
Influenza vaccine	Process	1,037	70.8%
Blood pressure ≤ 140/90	Intermediate Outcome	1,037	81.4%
Ischemic Vascular Disease			
Statin if indicated	Process	2,452	96.6%
Antiplatelet agent	Process	2,651	90.2%
Annual LDL	Process	2,672	87.4%
Beta-blocker if indicated	Process	2,611	73.3%
Pneumococcal vaccine	Process	2,672	71.4%
Influenza vaccine	Process	2,672	67.6%
Blood pressure ≤ 135/80	Intermediate Outcome	2,672	67.5%
Prevention			
Pneumococcal vaccine	Process	7,729	77.8%
Breast cancer screening	Process	14,983	75.9%
Cervical cancer screening	Process	34,992	75.8%

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Measures[*]	**Type of Measure**	**#of Eligible Patients**	**Percent Meeting Measure**
Diabetes			
Influenza vaccine	Process	14,622	59.9%

[*] See Table, Supplemental Digital Content 1, for a more detailed description of the measure criteria.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Measures evaluated at primary care physician (PCP) and PCP-group levels.

Measures*	Casemix- Adjusted PCP ICC† (SE)	Panel Size Needed for 80% Reliability‡	PCPs with Adequate Panel Size§	Casemix- Adjusted PCP Group ICC ∥ (SE)	Panel Needed for 80% Reliability‡	PCP Groups with Adequate Panel Size§
Diabetes						
ACE or ARB if proteinuric.	2.3 (2.2)	--	--	1.3 (1.3)	--	--
Statin if indicated	5.1 (1.8)¶	75	16.6%	0.8 (1.2)	--	--
Annual LDL	3.5 (1.2)¶	110	11.4%	0.7 (0.7)	--	--
Semi-annual PCP evaluation	2.7 (1.0)¶	145	4.0%	0.6 (0.5)	--	--
Semi-annual hemoglobin A1c	1.9 (0.8)¶	203	1.3%	1.5 (0.8)	--	--
Annual foot examination	7.1 (1.4)¶	53	39.6%	0.1 (0.5)	--	--
Pneumococcal vaccine	6.7 (1.4)¶	57	36.9%	3.0 (1.5)	--	--
Influenza vaccine	2.3 (0.7)¶	170	2.7%	0.5 (0.4)	--	--
Monitor for nephropathy	2.5 (1.2)¶	155	0%	0 (0)	--	--
Annual eye examination	4.7 (1.0)¶	81	20.8%	0.2 (0.4)	--	--
Antiplatelet agent if indicated	5.1 (1.1)¶	75	20.9%	0.7 (0.6)	--	--
Hemoglobin A1c ≤ 7%	1.0 (0.6)	--	--	0.2 (0.3)	--	--
Hemoglobin A1c ≤ 9%	0 (0)	--	--	0 (0)	--	--
Blood pressure ≤ 135/80	1.6 (0.5)¶	255	0%	0.6 (0.4)	--	--
Congestive Heart Failure						
Ejection fraction (EF) assessed	0 (0)	--	--	2.6 (1.8)	--	--
Annual creatinine and potassium if on ACE or ARB	5.3 (8.9)	--	--	0 (0)	--	--
Statin if indicated	6.5 (6.2)	--	--	0 (0)	--	--
ACE or ARB if EF< 40%	0.1 (9.3)	--	--	0 (0)	--	--
Beta-blocker if EF< 40%	0 (0)	--	--	0 (0)	--	--
Annual LDL	0 (0)	--	--	0 (0)	--	--
Pneumococcal vaccine	4.9 (4.0)	--	--	0.3 (1.5)	--	--
Influenza vaccine	1.9 (2.3)	--	--	0 (0)	--	--

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Measures[*]	Casemix- Adjusted PCP ICC[†] (SE)	Panel Size Needed for 80% Reliability[‡]	PCPs with Adequate Panel Size[§]	Casemix- Adjusted PCP Group ICC[‖] (SE)	Panel Needed for 80% Reliability[‡]	PCP Groups with Adequate Panel Size[§]
Blood pressure ≤ 140/90	3.1 (3.3)	--	--	0 (0)	--	--
Ischemic Vascular Disease						
Statin if indicated	0 (0)	--	--	1.8 (2.8)	--	--
Antiplatelet agent	1.0 (2.0)	--	--	1.5 (1.4)	--	--
Annual LDL	5.3 (2.3)[¶]	75	4.2%	0 (0)	--	--
Beta-blocker if indicated	0.9 (1.2)	--	--	1.2 (0.9)	--	--
Pneumococcal vaccine	2.1 (1.2)	--	--	1.6 (1.2)	--	--
Influenza vaccine	1.9 (1.3)	--	--	0.2 (0.6)	--	--
Blood pressure ≤ 135/80	3.1 (3.3)	--	--	0 (0)	--	--
Prevention						
Pneumococcal vaccine	4.6 (1.1)[¶]	83	16.9%	1.2 (0.9)	--	--
Breast cancer screening	7.8 (1.2)[¶]	48	60.1%	3.8 (1.6)[¶]	102	100
Cervical cancer screening	3.5 (0.6)[¶]	110	62.9%	2.8 (1.3)[¶]	139	94.1
Influenza vaccine	2.4 (0.5)[¶]	162	15.7%	0.8 (0.4)	--	--

[*] see Table, Supplemental Digital Content 1, for a more detailed description of the measure criteria

[†] The “Case-mix-adjusted PCP ICC” (primary care physician intra-class correlation coefficient) is the amount of total variance in the measure score that is due to differences between PCPs after controlling for patient factors, grouping by practice site, and random measurement error.

[‡] The panel size needed to obtain 80% reliability in a PCP or PCP Group's measure score was calculated using the case-mix-adjusted ICC and the Spearman-Brown Prophecy Formula.

[§] Percentage of PCPs or PCP Groups with an adequate panel size to achieve 80% reliability.

[‖] The “Case-mix-adjusted PCP Group ICC” is the amount of total variance in the measure score that is due to differences between PCPs grouped by practice site after controlling for patient factors and random measurement error.

[¶] Denotes statistically significant ICC.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 3

Single-disease and overall composite measures evaluated at primary care physician (PCP) level.

Composite Measure	Total Number of Measure Observations	Measure Adherence	Measures per PCP Mean (range)	Casemix- Adjusted PCP ICC[*] (SE)	Panel Needed for 80% Reliability[†]	PCPs with Adequate Panel Size[‡]
Diabetes	85,709	72.5%	575 (10-2,478)	1.9 (0.4)[§]	205	74.5%
CHF	4,967	81.6%	38 (3-168)	1.2 (0.9)	--	--
Ischemic Vascular Disease	15,949	76.2%	112 (5-641)	0.8 (0.4)[§]	530	2.1%
Prevention	72,326	72.8%	444 (2-1,877)	3.2 (0.5)[§]	120	76.7%
Overall	164,765	73.4%	1,011 (2-3,925)	5.4 (1.2)[§]	71	85.3%

[*] The "Case-mix-adjusted PCP ICC" (primary care physician intra-class correlation coefficient) is the amount of total variance in the composite score that is due to differences between PCPs after controlling for patient factors, measure factors, grouping by practice site, and random measurement error.

[†] The panel size needed to obtain 80% reliability in a PCP's score for a single-disease or overall composite using the case-mix-adjusted ICC and the Spearman-Brown Prophecy Formula.

[‡] Percent of PCPs with an adequate panel size to achieve 80% reliability for single-disease or overall composite measures.

[§] Denotes statistically significant ICCs.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 4

Single-disease and overall composites evaluated at primary care physician (PCP) group level.

Composite Measure	Total Number of Measure Observations	Measure Adherence	Measures per PCP Group Mean (range)	Casemix- Adjusted PCP Group ICC[*] (SE)	Panel Needed for 80% Reliability[†]	PCP Groups with Adequate Panel Size[‡]
Diabetes	85,709	71.6%	5,042 (1,493-10,420)	0.2 (0.2)	--	--
CHF	4,967	82.9%	292 (73-665)	0 (0)	--	--
Ischemic Vascular Disease	15,949	76.2%	938 (185-2,069)	0.2 (0.2)	--	--
Prevention	72,326	72.4%	4,255 (1,389-9,473)	1.0 (0.5)[§]	400	100%
Overall	164,765	73.8%	9,692 (2,877 – 20,802)	3.2 (1.6)[§]	121	100%

[*] The “Case-mix-adjusted PCP-Group ICC” (primary care physician-group intra-class correlation coefficient) is the amount of total variance in the composite score that is due to differences between PCPs grouped by site after controlling for patient factors, measure factors, and random measurement error.

[†] The panel size needed to obtain 80% reliability in a PCP Group's score for a single-disease or overall composite using the case-mix-adjusted ICC and the Spearman-Brown Prophecy Formula. Only reported for statistically significant case-mix-adjusted ICCs.

[‡] Percent of PCP Groups with an adequate panel size to achieve 80% reliability for single-disease or overall composite measures.

[§] Denotes statistically significant ICCs