The Reliability of Graduate Medical Education Quality of Care Clinical Performance Measures

Jung G. Kim, PhD, MPH
Hector P. Rodriguez, PhD, MPH
Eric S. Holmboe, MD
Kathryn M. McDonald, PhD, MM
Lindsay Mazotti, MD
Diane R. Rittenhouse, MD, MPH
Stephen M. Shortell, PhD, MBA, MPH
Michael H. Kanter, MD

ABSTRACT

Background Graduate medical education (GME) program leaders struggle to incorporate quality measures in the ambulatory care setting, leading to knowledge gaps on how to provide feedback to residents and programs. While nationally collected quality of care data are available, their reliability for individual resident learning and for GME program improvement is understudied.

Objective To examine the reliability of the Healthcare Effectiveness Data and Information Set (HEDIS) clinical performance measures in family medicine and internal medicine GME programs and to determine whether HEDIS measures can inform residents and their programs with their quality of care.

Methods From 2014 to 2017, we collected HEDIS measures from 566 residents in 8 family medicine and internal medicine programs under one sponsoring institution. Intraclass correlation was performed to establish patient sample sizes required for 0.70 and 0.80 reliability levels at the resident and program levels. Differences between the patient sample sizes required for reliable measurement and the actual patients cared for by residents were calculated.

Results The highest reliability levels for residents (0.88) and programs (0.98) were found for the most frequently available HEDIS measure, colorectal cancer screening. At the GME program level, 87.5% of HEDIS measures had sufficient sample sizes for reliable measurement at alpha 0.7 and 75.0% at alpha 0.8. Most resident level measurements were found to be less reliable.

Conclusions GME programs may reliably evaluate HEDIS performance pooled at the program level, but less so at the resident level due to patient volume.

Introduction

Residents training in ambulatory care settings are expected to enter unsupervised practice prepared to deliver high-quality patient care. Quality measures for appropriate prevention screening and evidence-based management of chronic conditions like diabetes and hypertension are essential to identify quality of care gaps and monitor quality improvement efforts, especially for patients subject to health disparities. While the importance of examining clinical performance in graduate medical education (GME) for outcomes-based measures, quality improvement, and program accountability has long been recognized, incorporating these measures into GME for resident assessment and program evaluation is still in question.[1-3]

Quality of care measures for physicians' patient health record audits and feedback create standards of accountability to the public and their payers, including GME's largest funder, the Centers for Medicare & Medicaid Services (CMS).[4,5] Rich sources of quality of care measures are available with CMS partnerships with the Core Quality Measures Collaborative and the National Committee for Quality Assurance (NCQA). While these partnerships are intended to establish standard quality of care measures for practicing physicians and their patients, the applicability of these publicly reported measures on individual resident and GME program performance in ambulatory care settings remains unclear.

Publicly reported quality of care data requires a sufficient number of patients within each quality measure to meet measurement standards for higher stakes purposes.[6-8] In ambulatory care–based GME, adequate patient volume is a critical component in preparing residents to become unsupervised practicing physicians. However, residency programs vary widely in their volume of patients and continuity of care.[9,10] It remains unclear whether the volume of patients in residency programs allows for resident- and program-level comparisons of quality of care and whether publicly reported quality measures are sufficiently reliable to inform assessments of residents and evaluate GME program performance across a sponsoring institution.[8,9,11] Prior studies estimating practicing physician group-level reliability of quality of care measures report sufficient patient sample sizes can be achieved when physicians are pooled collectively rather than assessed individually.[6,11,12]

DOI: http://dx.doi.org/10.4300/JGME-D-21-00706.1

This study examines the reliability of one set of publicly reported quality of care measures, the Healthcare Effectiveness Data and Information Set (HEDIS), and compares the clinical performance across family medicine and internal medicine GME residents and their programs within a sponsoring institution's health system. We estimated both the resident- and program-level reliability of publicly reported clinical performance measures in GME programs to determine whether patient volumes are sufficiently reliable to incorporate HEDIS measures to inform residents and their programs about their quality of care for improvement purposes.

Objectives
Examining the reliability of the Healthcare Effectiveness Data and Information Set (HEDIS) clinical performance measures in family medicine and internal medicine graduate medical education (GME) programs to determine whether HEDIS measures can inform programs with their quality of care.

Findings
At the GME program level, the majority of HEDIS measures had sufficient sample sizes for reliable measurement at the alpha 0.7 level, and most resident level measurements were found to be less reliable.

Limitations
Examining a single health system's use of HEDIS measures for GME programs with recognition that the capacity to access and critically examine quality of care measures is less available at other sponsoring institutions.

Bottom Line
GME programs should strive to access quality of care measures from their sponsoring institution to reliably evaluate the impact of patient volume on their residents' clinical performance and program quality improvement opportunities.

Methods

Study Sample

We studied a convenience sample of 566 resident physicians training over 3 years for all 8 accredited family medicine and internal medicine GME programs sponsored by the Kaiser Permanente Southern California health system between 2014 and 2018.

Measures

We analyzed the annual HEDIS performance associated with resident physicians for each calendar year between 2014 and 2017 by utilizing readily available HEDIS measures linked to residents' care that were collected by the health system's electronic health record and audited by the Kaiser Permanente Quality and Clinical Analysis division. HEDIS performance on our sample's patient panels are routinely collected as part of the health system's population health management and quality improvement efforts, and publicly reported annually to the NCQA.[13] Partnering with the Kaiser Permanente Department of Clinical Analysis, we extracted and examined 8 HEDIS measures based on the most frequently available measures for our sample's patient panel. These HEDIS measures are described under the NCQA's Effectiveness of Care category: diabetes management, prevention health screening (cancer), cardiovascular health (blood pressure control and cholesterol level management), and monitoring of patients on persistent medications.[8] Annual performance for each HEDIS measure was scored using the NCQA-defined HEDIS criteria of care that were sufficiently met by residents (numerator) divided by eligible patients assigned to a resident (denominator).[7] To characterize and benchmark our sample's HEDIS performance against national HEDIS mean performance, we obtained data from the publicly available NCQA website for the 2017 Commercial Health Plan reporting year and linked the national mean performance to each HEDIS measure.[14]

Program Characteristics

To characterize the Accreditation Council for Graduate Medical Education–accredited family medicine and internal medicine programs, residents with HEDIS performance data were linked to their GME program's data from the publicly available 2016-2017 American Medical Association's and Association of American Medical College's National GME Census.[15] Program characteristics include age and size of program, resident to faculty ratio, percent outpatient time for postgraduate year 1 (PGY-1) residents, and annual non-emergency department outpatient visits.

Statistical Analysis

Frequency, mean, and standard deviation were calculated to describe the program characteristics and program-level HEDIS scores. To estimate the reliability of each HEDIS measure, we calculated the intraclass correlation (ICC) at the resident and GME program levels. ICCs were calculated using unadjusted one-way random effects analysis of variance models, which estimated the variability attributed to HEDIS score differences within and between programs, based on the number of patients seen. The Spearman-Brown Prophecy Formula was used to estimate the patient sample sizes required for reliable resident- and program-level performance at the reliability (alpha) levels of 0.7 and 0.8.[16] These alpha levels are psychometrically accepted measurement standards for determining their respective unit-level reliability.[17] To evaluate whether sufficient samples were available, differences between the estimated patient sample sizes required were subtracted from

TABLE 1
Characteristics of 8 Graduate Medical Education Programs

Characteristic	No. (%)/Mean (SD)
Specialty	
Family medicine	6 (75)
Internal medicine	2 (25)
Age of program in years[a]	28.3 (17.5)
No. of residents/program	24.5 (9.6)
Resident-faculty ratio	1.2 (0.9)
% outpatient time for PGY-1 residents	31.5 (13.6)
Annual non-ED outpatient visits	104 569.9 (47 143.7)

Abbreviations: PGY, postgraduate year; ED, emergency department.
[a] Age of program is calculated by the last date of the study period (December 31, 2018) from the program's initial ACGME accreditation approval date.

the actual mean patient sample sizes available for the HEDIS measure.

Analyses were conducted using STATA 15.1 (StataCorp, College Station, TX) and Microsoft Excel (Microsoft Corporation, Redmond, WA).

This study was determined as non-human subjects research from the Kaiser Permanente Southern California Institutional Review Board.

Results

Table 1 describes the program characteristics for the sample of 6 family medicine and 2 internal medicine residencies (N=566 residents). The mean (SD) program age was 28.3 (14.5) years, with programs training a mean (SD) of 24.5 (9.6) residents/program and resident/faculty ratio of 1.2 (0.9). The mean (SD) reported proportion of training time in the outpatient setting for PGY-1s was 31.5% (13.6). The mean (SD) number of annual non-emergency department ambulatory care patient visits across all GME programs' medical centers was 104 569.9 (47 143.7).

Table 2 describes the 8 HEDIS quality measures examined. The availability of HEDIS measures per resident ranged from 32.7% to 97.2%, with colorectal cancer screening as the most frequent HEDIS measure reported (97.2%). The mean (SD) patient counts per resident for each HEDIS measure ranged from 13 (10) to 53 (45) patients. At the program level, 100% of HEDIS measures were available for all programs, with a mean (SD) patient counts per program for HEDIS measures ranging from 748 (258) to 5165 (1311) patients. The overall HEDIS score performance for our sample of programs was observed to be higher across all national HEDIS results as reported by the NCQA.

Table 3 summarizes the ICCs (95% confidence intervals) for HEDIS score variation and estimated reliability at the resident and program levels, based on the available number of patients for each HEDIS measure and performance year. Overall, the largest variation (ICC) for HEDIS measures at both the resident and program level were observed for hemoglobin A1c (HbA1c) <8.0 and ≤ 9.0 levels (0.059 and 0.057 for residents; 0.016 and 0.019 for programs). HEDIS score reliability increased when pooled at the program level. At the resident level, the estimated reliability levels ranged from 0.19 to 0.88, with colorectal cancer screening reporting the highest reliability level. At the program level, reliability ranged from 0.46 to 0.98, with colorectal cancer screening reporting the highest reliability (0.98), followed by cervical cancer screening (0.97). The lowest reliability

TABLE 2
Frequencies and Performance of 2014-2017 HEDIS Scores by 566 Residents Training in 8 Programs

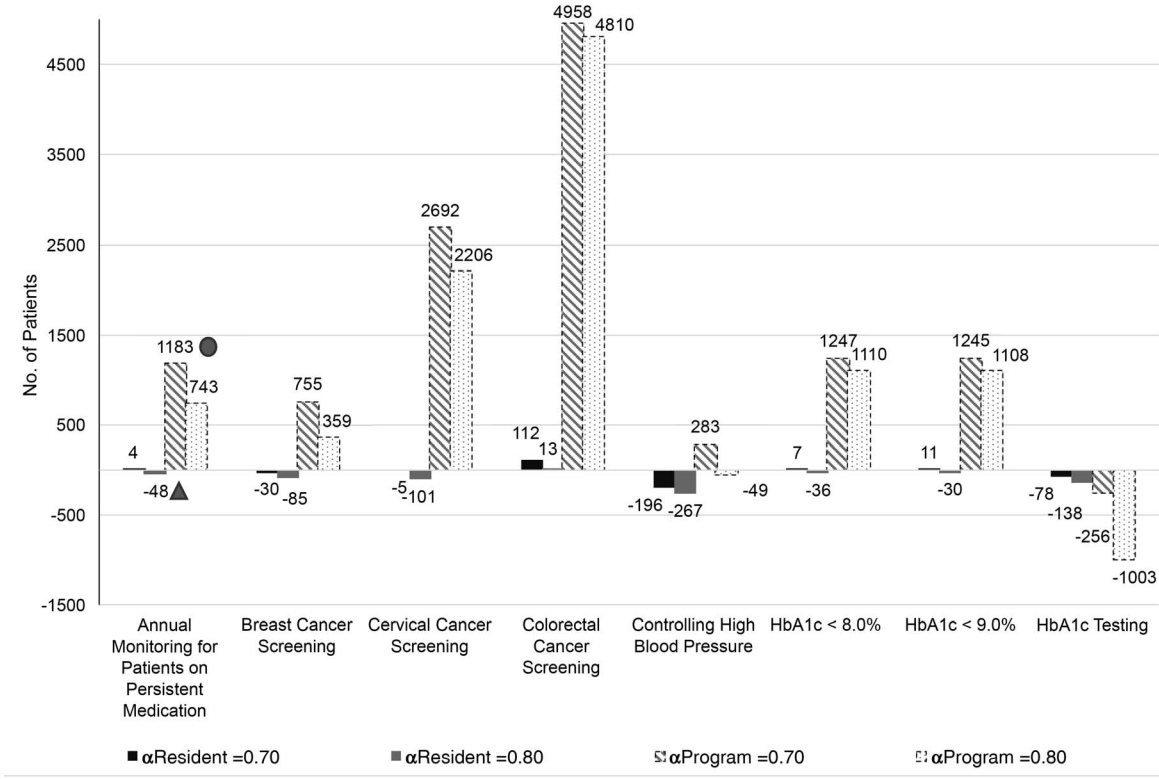
HEDIS Measure	No. (%) of Residents With Measure	Mean (SD) Patient Counts per Resident for Measure	No. (%) of Programs With Measure	Mean (SD) Patient Counts per Program for Measure	Overall Sample Mean (SD) HEDIS Performance Score	National Mean HEDIS Performance Score,[a] %
Annual monitoring for patients on persistent medication	536 (94.7)	29 (24)	8 (100)	1799 (273)	85.4 (15.2)	52.9
Breast cancer screening	517 (91.3)	22 (17)	8 (100)	1309 (230)	80.4 (17.6)	69.2
Cervical cancer screening	531 (93.8)	53 (45)	8 (100)	3372 (552)	81.9 (15.3)	68.9
Colorectal cancer screening	550 (97.2)	32 (31)	8 (100)	5165 (1311)	74.3 (18.4)	65.9
Controlling high blood pressure	539 (95.2)	14 (12)	8 (100)	748 (258)	82.3 (17.4)	63.2
HbA1c levels (<8.0)	185 (32.7)	13 (10)	8 (100)	1438 (436)	59.4 (21.4)	57.3
HbA1c levels (<9.0)	185 (32.7)	13 (10)	8 (100)	1438 (436)	92.9 (11.1)	67.7
HbA1c testing	343 (60.6)	13 (11)	8 (100)	790 (155)	92.5 (13.9)	91.1

Abbreviations: HEDIS, Healthcare Effectiveness Data and Information Set; SD, standard deviation.
[a] Source: National Committee for Quality Assurance Commercial Health Plan.

TABLE 3
Intraclass Correlation Coefficients and Reliability Estimates for Resident- and Program-Level Mean Performance Scores Using One-Way Analysis of Variance

HEDIS Measure	Year	2014		2015		2016		2017	
	Level of Analysis	Resident	Program	Resident	Program	Resident	Program	Resident	Program
Annual monitoring for patients on persistent medication	ICC (95% CI)	0.038 (0.028-0.048)	0.005 (0.000-0.013)	0.040 (0.027-0.052)	0.016 (0.000-0.037)	0.033 (0.021-0.044)	0.007 (0.000-0.016)	0.030 (0.021-0.040)	0.0014 (0.000-0.003)
	Reliability for mean scores	0.66	0.92	0.77	0.96	0.68	0.91	0.70	0.74
Breast cancer screening	ICC (95% CI)	0.023 (0.015-0.032)	0.005 (0.000-0.012)	0.032 (0.020-0.043)	0.002 (0.000-0.006)	0.034 (0.021-0.047)	0.004 (0.000-0.011)	0.025 (0.016-0.034)	0.0056 (0.000-0.012)
	Reliability for mean scores	0.47	0.87	0.66	0.77	0.61	0.82	0.60	0.89
Cervical cancer screening	ICC (95% CI)	0.018 (0.013-0.023)	0.010 (0.000-0.025)	0.019 (0.012-0.025)	0.007 (0.000-0.017)	0.016 (0.010-0.022)	0.0015 (0.000-0.003)	0.019 (0.013-0.024)	0.003 (0.000-0.007)
	Reliability for mean scores	0.63	0.97	0.73	0.96	0.63	0.80	0.74	0.92
Colorectal cancer screening	ICC (95% CI)	0.029 (0.022-0.035)	0.011 (0.000-0.026)	0.031 (0.022-0.040)	0.010 (0.000-0.023)	0.031 (0.022-0.041)	0.013 (0.000-0.029)	0.022 (0.016-0.028)	0.009 (0.000-0.021)
	Reliability for mean scores	0.73	0.97	0.88	0.98	0.85	0.98	0.85	0.98
Controlling high blood pressure	ICC (95% CI)	0.0093 (0.0018-0.016)	0.003 (0.000-0.007)	0.005 (0.00-0.011)	0.000 (0.000-0.001)	0.026 (0.013-0.038)	0.006 (0.000-0.014)	0.013 (0.005-0.020)	0.003 (0.000-0.008)
	Reliability for mean scores	0.19	0.75	0.17	0.76	0.46	0.81	0.33	0.76
HbA1c levels (<8.0)	ICC (95% CI)	0.031 (0.019-0.041)	0.013 (0.000-0.030)	0.051 (0.035-0.067)	0.016 (0.000-0.037)	0.059 (0.040-0.079)	0.012 (0.000-0.028)	0.047 (0.033-0.060)	0.0087 (0.000-0.019)
	Reliability for mean scores	0.45	0.92	0.79	0.96	0.78	0.94	0.78	0.94
HbA1c levels (<9.0)	ICC (95% CI)	0.041 (0.028-0.053)	0.019 (0.000-0.044)	0.052 (0.036-0.069)	0.017 (0.000-0.040)	0.057 (0.038-0.075)	0.010 (0.000-0.022)	0.043 (0.031-0.056)	0.007 (0.000-0.017)
	Reliability for mean scores	0.53	0.95	0.79	0.96	0.77	0.92	0.77	0.94
HbA1c testing	ICC (95% CI)	0.017 (0.0069-0.028)	0.009 (0.000-0.022)	0.021 (0.010-0.032)	0.003 (0.000-0.008)	0.019 (0.007-0.031)	0.000 (0.000-0.001)	0.026 (0.016-0.037)	0.000 (0.000-0.002)
	Reliability for mean scores	0.27	0.87	0.43	0.72	0.36	0.53	0.51	0.46

Abbreviations: HEDIS, Healthcare Effectiveness Data and Information Set; ICC, intraclass correlation; CI, confidence interval.



Differences are the delta between number of seen and estimated number of patients to meet specified alpha level
▲ *Negative values indicate insufficient number of patients to meet specified alpha level*
● *Positive values indicate sufficient and excess number of patients to meet specified alpha level*

FIGURE
Differences Between Number of Patients Seen and Estimated Number of Patients Needed for Resident- and Program-Level Reliability at 0.7 and 0.8 Alpha Levels for 2014-2017 HEDIS Measures

at the resident level were observed for controlling high blood pressure, with a range of 0.17 to 0.33.

The FIGURE illustrates the differences between the number of patients seen by residents and the estimated number of patients needed for resident- and program-level reliability, at the 0.7 and 0.8 alpha levels. Positive values indicate the HEDIS measure is reliably sufficient at the specified alpha reliability levels, whereas negative values indicate the HEDIS measure has insufficient patient counts. At the program level, 7 of 8 measures (87.5%) were reliable at alpha=0.7 and 6 of 8 measures (75.0%) at alpha=0.8, with the exception of HbA1c testing. At the resident level, 50% of HEDIS measures had sufficient number of patients to be reliable at the alpha 0.7 level for and 12.5% at an alpha level 0.8. Colorectal cancer screening was the sole HEDIS measure meeting both alpha levels at the resident level.

Discussion

This study found that, within a single sponsoring institution's family medicine and internal medicine residency programs, the minimum level of reliability was met for 7 of 8 HEDIS measures examined at the program level, but only half of the measures at the resident level. Our findings are consistent with other studies of practicing physicians that found less reliability at the individual level versus the practice level and highlights the following key takeaways: (1) programs may be able to reliably examine the quality of care for their residents when HEDIS measures are pooled by program, and (2) that patient volume impacts the reliability of residents' quality of care performance. These findings contribute to the argument that examining available quality measures, like HEDIS, is needed to understand their utility when assessing resident performance and evaluating program improvement efforts.[3]

Resident quality of care performance for meeting colorectal cancer screening had the highest reliability estimates, which was the most frequently available HEDIS measure for residents (97.2%). In contrast, the hypertension control measure had insufficient patient volume and was found to be the least reliable. This could be due to colorectal cancer screening tests

that can be ordered and completed after one visit, versus hypertension control efforts that are more complex and require multiple follow-up visits and associated laboratory testing. These findings have the following implications for programs: (1) quality of measures with high reliability may be a good target for program directors to incorporate as a proxy for quality of care performance and quality improvement efforts, at least at the program level, and (2) quality of care measures should be available for essentially all residents within a program and have sufficient patient volume. The latter highlights the impact of the number of patients cared for by residents and how much time residents are scheduled in ambulatory care when measuring for their quality of care to achieve core physician competencies like Systems-Based Practice. Prior studies on GME accreditation issues found that family medicine programs struggle with scheduling resident continuity clinic time, and the current national average of 25% of ambulatory care time for PGY-1 residents training in family medicine and internal medicine may be insufficient.[10,18] Hence, programs should consider the impact of resident scheduling on the opportunity to care for a sufficient number of patients in the ambulatory care settings and by extension enhance the reliability of quality of care measures to support their development of core physician competencies.

The generalizability of our findings is limited given the focus on a single sponsoring institution's health system and its use of HEDIS measures. We also recognize that other sponsoring institutions may not have the current capacity to extract and examine quality of care measures specific to their residency programs. However, at the program level, the ability to examine patient care measures, including patient volume, is an important component of residency review committee reporting requirements.[19,20] There is also the need for health systems' quality of care entities to partner with their GME programs to provide access to quality measures in pursuing shared interests in improving the quality of care. This aligns with the common program requirements that call for programs to engage in quality improvement efforts, that residents routinely receive quality of care data related to their patients, and the recent harmonized Milestones, specifically in the Practice-Based Learning and Improvement competency, that calls for evidence-based and informed practice for performance improvement.[21,22]

We also recommend that GME programs examine the type and context of clinical performance measures for residents and programs prior to incorporating them for assessment and program evaluation purposes. Additionally, programs and their sponsoring institutions needing to meet accreditation and public national quality of care reporting standards should consider reporting pooled ambulatory care-based clinical performance measures at the program level. Pooled program-level data may also help identify quality of care gaps and the monitoring of quality improvement efforts both within sponsoring institutions' health systems and across programs nationally.

Future research should continue to identify best practices that increase the capacity to access and critically examine currently available quality of care measures on GME programs. Additional studies with larger sample sizes will facilitate analysis of resident-level characteristics to better model and understand the ICCs for HEDIS at both the resident level and program level simultaneously. This could establish how much of the variation in quality of care for patients is attributed to individual resident characteristics versus health system–level factors, which can assist in targeting improvement interventions. It also remains unclear how limited face-to-face patient access and resident training time disruption has affected clinical performance due to unprecedented events such as the COVID-19 pandemic, which may impact the reliability of GME quality of care measurements. Further research should examine whether GME-related quality of care performance has been impacted by reduced face-to-face patient volume and the increase of other forms of patient-physician encounters.

Conclusions

Examining HEDIS measures to compare and evaluate GME program performance had sufficient reliability at the program level, but was less reliable at the resident level. When available, HEDIS measures are one source that GME programs may find useful for program comparisons and informing the quality of care for their program's improvement efforts.

References

1. Edwards ST, Kim H, Shull S, Hooker ER, Niederhausen M, Tuepker A. Quality of outpatient care with internal medicine residents vs attending physicians in Veterans Affairs primary care clinics. *JAMA Intern Med*. 2019;179(5):711-713. doi:10.1001/jamainternmed.2018.8624
2. Asch DA, Nicholson S, Srinivas S, Herrin J, Epstein AJ. Evaluating obstetrical residency programs using patient outcomes. *JAMA*. 2009;302(12):1277-1283. doi:10.1001/jama.2009.1356
3. Smirnova A, Sebok-Syer SS, Chahine S, et al. Defining and adopting clinical performance measures in graduate

medical education: where are we now and where are we going? *Acad Med*. 2019;94(5):671-677. doi:10.1097/ACM.0000000000002620
4. Martin P, Zindel M, Nass S, eds. *Graduate Medical Education Outcomes and Metrics: Proceedings of a Workshop*. Washington, DC: National Academies Press; 2018. doi:10.17226/25003
5. Wong BM, Baum KD, Headrick LA, et al. Building the bridge to quality: an urgent call to integrate quality improvement and patient safety education with clinical care. *Acad Med*. 2020;95(1):59-68. doi:10.1097/ACM.0000000000002937
6. Rodriguez HP, von Glahn T, Chang H, Rogers WH, Safran DG. Measuring patients' experiences with individual specialist physicians and their practices. *Am J Med Qual*. 2009;24(1):35-44. doi:10.1177/1062860608326418
7. National Committee for Quality Assurance. HEDIS Measures and Technical Resources. NCQA. Published 2020. Accessed November 15, 2020. https://www.ncqa.org/hedis/measures/
8. National Committee for Quality Assurance. HEDIS & Quality Measurement, HEDIS Measures. Accessed November 15, 2020. http://www.ncqa.org/HEDISQualityMeasurement/HEDISMeasures/HEDIS2015.aspx
9. Walker J, Payne B, Clemans-Taylor BL, Snyder ED. Continuity of care in resident outpatient clinics: a scoping review of the literature. *J Grad Med Educ*. 2018;10(1):16-25. doi:10.4300/JGME-D-17-00256.1
10. Pugno PA, Epperly TD. Residency Review Committee for Family Medicine: an analysis of program citations. *Fam Med*. 2005;37(3):174-177.
11. Sequist TD, Schneider EC, Li A, Rogers WH, Safran DG. Reliability of medical group and physician performance measurement in the primary care setting. *Med Care*. 2011;49(2):126-131. doi:10.1097/MLR.0b013e3181d5690f
12. Scholle SH, Roski J, Dunn DL, et al. Availability of data for measuring physician quality performance. *Am J Manag Care*. 2009;15(1):67-72.
13. Kanter MH, Lindsay G, Bellows J, Chase A. Complete care at Kaiser Permanente: transforming chronic and preventive care. *Jt Comm J Qual Patient Saf*. 2013;39(11):484-494. doi:10.1016/s1553-7250(13)39064-3
14. National Committee for Quality Assurance. HEDIS Measures and Technical Resources: Cervical Cancer Screening. NCQA. Accessed July 27, 2019. https://www.ncqa.org/hedis/measures/cervical-cancer-screening/
15. American Medical Association. FREIDA Residency Program Database. Accessed February 12, 2019. https://freida.ama-assn.org
16. Caci HM. SBROWNI: Stata Module to Calculate Spearman-Brown Reliability Correction for Test Length. Boston College Department of Economics; 1998. Accessed June 24, 2019. https://ideas.repec.org/c/boc/bocode/s351002.html
17. Nunnally JC, Bernstein IH. *Psychometric Theory*. New York, NY: McGraw-Hill; 1994.
18. Kim JG, Rodriguez HP, Shortell SM, Fuller B, Holmboe ES, Rittenhouse DR. Factors associated with family medicine and internal medicine first-year residents' ambulatory care training time. *Acad Med*. 2021;96(3):433-440. doi:10.1097/ACM.0000000000003522
19. Accreditation Council for Graduate Medical Education. ACGME Program Requirements for Graduate Medical Education in Internal Medicine. Accessed March 7, 2022. https://www.acgme.org/globalassets/pfassets/programrequirements/140_internalmedicine_2020.pdf
20. Accreditation Council for Graduate Medical Education. ACGME Program Requirements for Graduate Medical Education in Family Medicine. Accessed March 7, 2022. https://www.acgme.org/globalassets/PFAssets/ProgramRequirements/120_FamilyMedicine_2020.pdf
21. ACGME Common Program Requirements (Residency). Accessed March 7, 2022. https://www.acgme.org/globalassets/PFAssets/ProgramRequirements/CPRResidency2021.pdf
22. Edgar L, Roberts S, Holmboe E. Milestones 2.0: a step forward. *J Grad Med Educ*. 2018;10(3):367-369. doi:10.4300/JGME-D-18-00372.1

Jung G. Kim, PhD, MPH, is Assistant Professor, Kaiser Permanente Bernard J. Tyson School of Medicine, Department of Health Systems Science; **Hector P. Rodriguez, PhD, MPH,** is the Kaiser Permanente Professor of Health Policy and Management, University of California, Berkeley School of Public Health; **Eric S. Holmboe, MD,** is Chief Research, Milestone Development, and Evaluation Officer, Accreditation Council for Graduate Medical Education; **Kathryn M. McDonald, PhD, MM,** is the Bloomberg Distinguished Professor of Health Systems, Quality, and Safety, Johns Hopkins Schools of Medicine and Nursing; **Lindsay Mazotti, MD,** is Assistant Physician-in-Chief, Kaiser Permanente East Bay and Director, Clinical Experience/Associate Professor of Clinical Science, Kaiser Permanente School of Medicine; **Diane R. Rittenhouse, MD, MPH,** is Senior Fellow, Mathematica, and Professor, University of California, San Francisco; **Stephen M. Shortell, PhD, MBA, MPH,** is Blue Cross of California Distinguished Professor of Health Policy and Management Emeritus, Dean Emeritus, and Professor, Graduate School, University of California, Berkeley School of Public Health; and **Michael H. Kanter, MD,** is Chair and Professor of Clinical Science, Kaiser Permanente School of Medicine.

Funding: The authors report no external funding source for this study.

Conflict of interest: Dr Holmboe is an employee of the Accreditation Council for Graduate Medical Education and has written a clinical competency textbook published by Elsevier.

This work was previously presented at the virtual AcademyHealth Annual Research Meeting, July 28-August 6, 2020.

The authors would like to thank the Kaiser Permanente Southern California Department of Clinical Analysis and the Kaiser Permanente Program Directors of Internal Medicine committee. The authors also especially thank the patients of Kaiser Permanente and their partnership with us to improve their health.

Corresponding author: Jung G. Kim, PhD, MPH, Kaiser Permanente Bernard J. Tyson School of Medicine, kim.jung@kp.org, Twitter @jkay206

Received July 12, 2021; revisions received October 26, 2021, and February 23, 2022; accepted February 28, 2022.