Directly Observed Care: Can Unannounced Standardized Patients Address a Gap in Performance Measurement?

Saul J. Weiner, MD[1,2,3] *and Alan Schwartz, PhD*[3]

[1]VA Center of Innovation for Complex Chronic Healthcare, Jesse Brown VA Medical Center, Chicago, IL, USA; [2]Departments of Medicine and Pediatrics, University of Illinois at Chicago, Chicago, IL, USA; [3]Department of Medical Education, University of Illinois at Chicago, Chicago, IL, USA.

There are three potential sources of information for evaluating a clinician's performance: documentation, patient report, and directly observed care. Current measures draw on just two of these: data recorded in the medical record and surveys of patients. Neither captures an array of performance characteristics, including clinician attention to symptoms and signs while taking a history or conducting a physical exam, accurate recording in the medical record of information obtained during the encounter, evidence based communication strategies for preventive care counseling, and effective communication behavior. Unannounced Standardized Patients (USPs) have been widely deployed as a research strategy for systematically uncovering significant performance deficits in each of these areas, but have not been adopted for quality improvement. Likely obstacles include concerns about the ethics of sending health professionals sham patients, the technical challenges of the subterfuge, and concerns about the relatively small sample sizes and substantial costs involved. However, the high frequency of significant and remediable performance deficits unmasked by USPs, and the potential to adapt registration and record keeping systems to accommodate their visits, suggest that their selective and purposeful deployment could be a cost effective and powerful strategy for addressing a gap in performance measurement.

J Gen Intern Med 29(8):1183–7
DOI: 10.1007/s11606-014-2860-7

Clinician and health system performance is assessed by employing performance measures, a "set of technical specifications that define how to calculate a rate for some important indicator of quality."[1] The data are collected from two sources: records generated during the care delivery process, and surveys of patients. The former draws on information recorded in the medical record, such as a diagnosis or an order placed for a clinical test or treatment. The latter draws on information the patient provides about their experience of care. Over 90 % of health plans use the Healthcare Effectiveness Data and Information Set (HEDIS), which now includes the Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey, for performance measurement.[1]

Neither strategy collects data on the performance of healthcare professionals and staff during patient interactions. For instance, there are HEDIS measures on management of blood pressure, but no measures of whether blood pressure was measured correctly. There are measures of whether patients were reportedly advised to quit smoking, but not measures of whether motivational interviewing strategies were employed or whether the advising actually occurred. There are, in fact, an array of performance characteristics that require directly observing care to ascertain whether information recorded in the medical record is an accurate representation of care and of the health needs of the patient. Currently utilized performance measures rely on that record.

Unfortunately there is evidence that the medical record is not an accurate record of either patients' health care needs or of the care they have received. The evidence has accrued from studies employing unannounced standardized patients (USPs) who directly observe care and have been shown to be valid and reliable reporters of physician practice.[2] USPs are portrayed by individuals trained to role-play specific scripts in the clinical setting and then document what they observe. While not the only way to directly observe care—clinical encounters can be observed and assessed by colleagues, for instance—USPs have been described as the "gold standard" of clinician performance measurement because they are standardized, meaning they portray the same problem by the same patient across multiple encounters, allowing apples-to-apples comparisons of how different providers and health systems respond to a particular clinical scenario, they record (using checklists and sometimes audio) what occurs during an episode of care, and finally, they do so incognito, so that what they see represents "usual care," rather than best behavior .[2] Details of how USPs are trained, how scripts are developed, and the processes for monitoring their performance to assure they function effectively and safely in the clinical setting have been previously described.[3]

Received December 7, 2013
Revised March 10, 2014
Accepted March 27, 2014
Published online April 23, 2014

USP studies have called into question the reliability and adequacy of the medical record as a record of care delivery, and revealed a number of performance deficits not captured with current performance measures. They have documented both physicians recording physical exam maneuvers not conducted and failing to record preventive care services actually delivered. In one study, 33 % of physical exam maneuvers recorded in the medical record were never conducted—an error of commission,[4] while in another, 16 % of preventive care delivered was not recorded—an error of omission.[5] USPs portraying cases have documented less than 50 % of physicians performing fundoscopy or examining feet in diabetics,[6] a third of primary care physicians providing no skin protective counseling—not even recommendations to use a sunscreen—during a pre-work physical for an 18-year-old fair-skinned woman starting as a life guard at the beach,[7] and 40 % of rheumatologists missing the diagnosis of psoriatic arthritis because of a failure to examine the skin.[8] In a study we conducted, a third of general internists across a dozen primary care practices neglected to consider hypothyroidism in a middle-aged woman presenting for pre-operative evaluation for hip surgery who complained of constipation, weight gain, heavy menses and poor sleep.[9]

Despite the insight it provides, directly observed care is not a required component of quality assessment by any insurer or health plan of which we are aware. Directly observed care may be regarded as the missing component of a three-part strategy for the evaluation of quality and performance in health systems and individual encounters, as outlined in Table 1.

CHALLENGES OF EMPLOYING USPS FOR PERFORMANCE MEASUREMENT

Four questions surface in discussions about covertly and directly observing care: Is it ethical to trick health care providers into thinking they are seeing real patients for the purpose of evaluating their performance? Are the technical challenges of creating the subterfuge surmountable? Are sample sizes adequate to draw meaningful inferences from the data obtained? And, is this form of performance assessment affordable and cost-effective as compared with other forms?

Ethics. Lower stakes industries have long embraced the “secret shopper” as a quality assessment and improvement strategy, including large sectors of the retail industry. Secret shoppers are not standardized to reliably exhibit specific behaviors designed to evaluate respondents, so they do not bring the rigor of USPs to measurement, but the principle of periodic incognito observation as both a tool to evaluate quality and an incentive to improve performance is widely accepted outside of health care.[10] Organized medicine has been more reluctant. In 2008 the Council on Ethics and Judicial Affairs (CEJA) of the American Medical Association (AMA) recommended the use of secret shoppers under limited conditions, writing “Physicians have an ethical responsibility to engage in activities that contribute to continual improvements in patient care. One method for promoting such quality improvement is through the use of secret shopper ‘patients’ who have been appropriately trained to provide feedback about physician performance in the clinical setting,” but the AMA House of Delegates tabled the resolution and CEJA withdrew it after a significant opposition, including concerns that it represented a failure to view the physician as a professional.[11] In our large study employing USPs, seven institutional review board (IRB) committees representing each of the sites independently approved the protocol.[9] We argue for the legitimacy of employing USPs, given the significance and implications for patient care of the information that is revealed when care is systematically and directly observed, as demonstrated by the studies that have used this methodology of performance assessment.

Technical Challenges. Table 2 lists five criteria for optimally deploying USPs as a performance measure. Perhaps most challenging is the first, which entails covertly introducing a sham patient into a real practice—i.e. the subterfuge.[12] Actors who are chosen for USP work must demonstrate that they can consistently

Table 1. Directly Observed Care Employing USPs Completes a Three-Part Strategy of Performance Assessment

What is assessed	How measured	What it Does	Limitations
Care processes as recorded in the chart	Information extracted from medical record and claims data	Rates evidence based indicators of quality	Relies on accuracy and completeness of information in the medical record
Care processes as experienced by patients	Patient ratings using surveys	Captures how patients describe and rate their experiences receiving care	Self-selection bias (dissatisfied patients don’t return) and limited frame of reference (most patients don’t have extensive comparative experience)
Care processes as directly observed	Currently not measured. Unannounced standardized patients	Rates how staff and clinicians conduct procedures, elicit and process information from patients, and attend to their needs	Relatively small sample sizes, technical challenge of creating subterfuge, costs of USPs, opportunity cost of displacing a patient, coding costs

Table 2. Conditions for Employing USPs for Optimal Performance Assessment

- Clinicians are unaware they are seeing a USP
- Actors adhere to their scripts consistently
- Performance measures are based on research evidence
- Cases are constructed to assess behaviors that cannot be reliably assessed using simpler methods
- The overall number and pattern of cases is customized to answer important questions for stakeholders to the evaluation

portray a specific set of symptoms and personality traits while also adapting their character to real-life situations. At each site, typically one mid-level practice administrator who is authorized to register new patients and who supervises front desk staff schedules each visit. The USP may present as self-pay or is pre-registered in the electronic health record (EHR) as insured, so that when they arrive they are approved for the visit. Following the visit, the physician is typically notified once they have completed their note that the patient was a USP. Detection rates, which vary widely by study, are typically measured by asking physicians how suspicious they were that the patient they saw was a USP. One large study included a detailed analysis of "meaningful detection," which occurs when a physician is able to identify the USP from among patients seen in the prior two weeks and becomes suspicious either before or during the visit. Meaningful detection, significant because it could threaten the validity of data collected (largely by giving physicians a chance to perform better than usual), occurred 12.8 % of the time.[12] Following each encounter, the medical record is removed from the production environment after the physician's note is extracted, and orders are cancelled.

The technical complexities could be simplified and greater versatility introduced if systems were adapted specifically to accommodate USPs. Consider, for instance, an EHR with a design function for creating a simulated chart that could be pre-populated with demographics, notes, laboratory information and orders—all from a drop-down menu and library of options—that expands the repertoire of clinical scenarios used to assess clinicians and other members of the health care team. On the back end, the EHR system would code these charts as simulated and ephemeral, but to the user they would be indistinguishable from an authentic record. Similar tools have been in place for educational purposes for a decade, but these simulated medical records function outside of the production environment.[13]

Sample Size. Performance measures, even at the level of individual provider assessment, typically rely on a substantial sample size, such as the number of eligible patients who received a colonoscopy in a panel, or the proportion of diabetics who achieved an Hgb A1c in a target range. Standardized patients, however, have demonstrated evidence for validity sufficient for high stakes licensure exams with small samples sizes of 10–12 encounters.[14] A subscale measure of communication behavior based on only three SP encounters during a national clinical skills assessment exam predicted future complaints to regulatory authorities.[15] The predictive power may be related to the volume of relevant information collected from even a single visit. For instance, in one audio-recorded USP encounter, a clinician is heard repeatedly answering phone calls while the patient is attempting to disclose sensitive information, and misses four out of four clues embedded in the script of a significant underlying condition. The information value of a single USP visit could be expanded further with an analysis of the entire episode of care, starting from registration at the front desk, and inclusive of the waiting time, pre-visit assessment by a nurse or nursing assistant, and the exit process following the clinical encounter. Although USPs have not been substantially deployed outside of the clinical encounter in research protocols, "mystery patients" trained to focus on customer service are currently marketed by several companies. With appropriate training USPs may widen their lens to document, for instance, whether blood pressure is measured correctly pre-encounter, HIPAA compliant practices are adhered to, or screening questions are in fact asked rather than just recorded as asked. With this approach, the entire practice rather than the individual provider becomes the unit of analysis.

Cost. The major costs are the actors' time and project management. There is little published data on costs. In our 2008–2009 study employing eight actors, 399 USP visits were completed on a budget of $146,983, or $365/encounter, which included case construction, project management, costs of actor recruitment, training, and employment in the field, travel expenses, monitoring of role portrayal fidelity and checklist completion accuracy, data analysis and report generation. Hence, in 2014 adjusted dollars, we estimate USPs can provide a 20–30 encounter assessment of a clinical setting across the full experience of care, from the initial phone contact to customer service at check in, wait time, the skill of the pre-visit assessment, the visit itself, and concordance among visits, clinical documentation, and billing for $8–12 K. In addition, there is the opportunity cost of not seeing a real patient. In non-research contexts and with additional experience with the method at a particular site, costs are likely to decrease. A sufficient number of visits to a particular provider may yield reliable practitioner specific performance data. For instance, a high-stakes clinical skills assessment required for medical licensure has employed ten SP encounters with generalizability coefficients of 0.70–0.90.[16] USP costs should also be compared to the potential savings of identifying and correcting otherwise undetectable error prone behaviors that lead to inappropriate tests and treatments. The substantial cost of errors detectable only by directly observing care has been calculated in research employing USPs.[17] Finally, the modest expense for actionable information should be compared with other modalities that are also labor intensive, such as nurse audits of the medical record.

CONCLUSIONS

Directly observed care is arguably the missing piece in performance measurement at both the individual provider and practice systems levels, with the USP as a research tested strategy for systematically addressing the gap. There are several reasons that physicians and practices should embrace USPs. USPs have turned up evidence of significant, common, and costly deficits in care delivery that are currently unmeasured. Each USP, portraying a script customized to assess pre-selected components of interest, can systematically collect data for analysis of elements of customer service (e.g. telephone inquiry, way finding, handling difficult clients), provider performance (e.g. hand hygiene, communication behavior, preventive care counseling, critical elements of history taking and the physical exam), and documentation (reconciliation of what occurred during a visit with what was recorded as having occurred). A dozen USP visits, with appropriately selected cases, can meet high standards of evidence for validity and reliability for evaluating a clinical process or a particular provider. Participation in a USP program might count towards maintenance of certification and/or for performance improvement continuing medical education. USPs could also be deployed to provide in vivo assessment of a provider needing remediation. Finally, clinical practices aspiring to differentiate themselves as exceptional providers may wish to participate in a USP program to identify and address performance deficits through cycles of continuous quality improvement until all staff are consistently high performers when systematically, covertly and directly observed.

Several challenges confront adoption of USPs: First, registering and processing sham patients is a technical challenge, but could be simplified by adding features to the EHR for establishing simulated medical records. Second, there is resistance from organized medicine to endorse USPs. Such resistance is consistent with a history of physician skepticism about new strategies of performance assessment and quality improvement.[18] Third—and addressing this one will likely also resolve the second—is that payers are not yet calling for information about observed care. Performance measures take hold when payers, including Medicare, recognize their value and require them. The evidence suggests that certain critical measures of quality require systematic observation of care processes and interactions, and that doing so is feasible. We submit that further attention to this missing piece in assessing health care delivery is an important next step in evaluating and improving quality.

Funding/Support: *Drs. Weiner and Schwartz's research on unannounced standardized patients has been supported by the Veterans Affairs, Health Services Research and Development, and the National Board of Medical Examiners.*

Disclaimer: *The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the U.S. Department of Veterans Affairs or the United States government.*

Disclosures: *The authors are co-principals of the Institute for Practice and Provider Performance Improvement, founded to employ unannounced standardized patient assessments as a quality improvement service. They have not at this point received any fees or contracts. There are no other relationships or activities that could appear to have influenced the submitted work.*

Corresponding Author: *Saul J. Weiner, MD; University of Illinois at Chicago, 601 South Morgan Street, 2732 UH, Chicago, IL 60607, USA (e-mail: sweiner@uic.edu).*

REFERENCES

1. National Center for Quality Assurance. 2013, at http://www.ncqa.org/HEDISQualityMeasurement/PerformanceMeasurement.aspx. Accessed March 2014.
2. **Luck J, Peabody JW.** Using standardised patients to measure physicians' practice: validation study using audio recordings. BMJ. 2002;325:679.
3. **Glassman PA, Luck J, O'Gara EM, Peabody JW.** Using standardized patients to measure quality: evidence from the literature and a prospective study. Jt Comm J Qual Improv. 2000;26:644–53.
4. **Dresselhaus TR, Luck J, Peabody JW.** The ethical problem of false positives: a prospective evaluation of physician reporting in the medical record. J Med Ethics. 2002;28:291–4.
5. **Dresselhaus TR, Peabody JW, Lee M, Wang MM, Luck J.** Measuring compliance with preventive care guidelines: standardized patients, clinical vignettes, and the medical record. J Gen Intern Med. 2000;15:782–8.
6. **Krane NK, Anderson D, Lazarus CJ, et al.** Physician practice behavior and practice guidelines: using unannounced standardized patients to gather data. J Gen Intern Med. 2009;24:53–6.
7. **Hornung RL, Hansen LA, Sharp LK, Poorsattar SP, Lipsky MS.** Skin cancer prevention in the primary care setting: assessment using a standardized patient. Pediatr Dermatol. 2007;24:108–12.
8. **Gorter S, van der Heijde DM, van der Linden S, et al.** Psoriatic arthritis: performance of rheumatologists in daily practice. Ann Rheum Dis. 2002;61:219–24.
9. **Weiner SJ, Schwartz A, Weaver F, et al.** Contextual errors and failures in individualizing patient care: a multicenter study. Ann Intern Med. 2010;153:69–75.
10. **Finn A.** Shopper benchmarking of durable-goods chains and stores. J Serv Res. 2001;3:310–20.
11. **Levine M.** CEJA Report 3-A-08 Secret Shopper "Patients.". Chicago, IL: American Medical Association; 2008.
12. **Franz CE, Epstein R, Miller KN, et al.** Caught in the act? Prevalence, predictors, and consequences of physician detection of unannounced standardized patients. Health Serv Res. 2006;41:2290–302.
13. **Speedie SM, Niewoehner C.** The Minnesota Virtual Clinic: using a simulated EMR to teach medical students basic science and clinical

concepts. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium 2003:1013.

14. **van Zanten M, Boulet JR, McKinley D.** Using standardized patients to assess the interpersonal skills of physicians: six years' experience with a high-stakes certification examination. Health Commun. 2007;22:195–205.
15. **Tamblyn R, Abrahamowicz M, Dauphinee D, et al.** Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA. 2007;298:993–1001.
16. **Whelan GP, Boulet JR, McKinley DW, et al.** Scoring standardized patient examinations: lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA). Med Teach. 2005;27:200–6.
17. **Schwartz A, Weiner SJ, Weaver F, et al.** Uncharted territory: measuring costs of diagnostic errors outside the medical record. BMJ Qual Saf 2012.
18. **Audet AM, Doty MM, Shamasdin J, Schoenbaum SC.** Measure, learn, and improve: physicians' involvement in quality improvement. Health Aff (Millwood). 2005;24:843–53.