OPEN

Expert Consensus on Currently Accepted Measures of Harm

Merranda S. Logan, MD, MPH,†‡ Laura C. Myers, MD, MPH,†‡§ Hojjat Salmasian, MD, MPH, PhD,‡‖ David Michael Levine, MD, MPH, MA,‡¶ Christopher G. Roy, MD, MPH,‡** Mark E. Reynolds, BA,†† Luke Sato, MD,‡†† Carol Keohane, MS, RN,†† Michelle L. Frits, BA,¶ Lynn A. Volk, MHS,‡‡ Ruth N. Akindele, MBBS, MPH,¶ Juliette M. Randazza, BA,¶ Sevan M. Dulgarian, BS, BA,¶ David M. Shahian, MD,†‡§§ David Westfall Bates, MD, MSc,‡¶‡‡‖‖ and Elizabeth Mort, MD, MPH†¶¶****

Background: Twenty-five years after the seminal work of the Harvard Medical Practice Study, the numbers and specific types of health care measures of harm have evolved and expanded. Using the World Café method to derive expert consensus, we sought to generate a contemporary list of triggers and adverse event measures that could be used for chart review to determine the current incidence of inpatient and outpatient adverse events.

Methods: We held a modified World Café event in March 2018, during which content experts were divided into 10 tables by clinical domain. After a focused discussion of a prepopulated list of literature-based triggers and measures relevant to that domain, they were asked to rate each measure on clinical importance and suitability for chart review and electronic extraction (very low, low, medium, high, very high).

Results: Seventy-one experts from 9 diverse institutions attended (primary acceptance rate, 72%). Of 525 total triggers and measures, 67% of 391 measures and 46% of 134 triggers were deemed to have high or very high clinical importance. For those triggers and measures with high or very high clinical importance, 218 overall were deemed to be highly amenable to chart review and 198 overall were deemed to be suitable for electronic surveillance.

Conclusions: The World Café method effectively prioritized measures/triggers of high clinical importance including those that can be used in chart review, which is considered the gold standard. A future goal is to validate these measures using electronic surveillance mechanisms to decrease the need for chart review.

Key Words: patient safety, implementation, patient safety culture, coincidence analysis, primary care, patient-centered medical home

(*J Patient Saf* 2021;17: e1726–e1731)

Thousands of patients each year experience health care–related adverse events.[1–4] It is important that we understand the frequency of specific events and their variation among providers, in order to support improvement work aimed at reducing patient harm. Twenty-five years after the seminal work of the Harvard Medical Practice Study, as health care has evolved, so have the numbers and types of patient harms being tracked and investigated.[5,6] Common methods for adverse event measurement include voluntary reporting systems, trigger tools, administrative data analysis, manual chart review, clinical registries, patient complaints, malpractice claims analysis, and electronic data extraction.[7,8] Each has advantages and disadvantages, and use of methods is not standardized across the health care industry. For example, voluntary reporting systems represent a confidential vehicle to capture adverse events, but only a small percentage, typically 5%–20%, of adverse events are reported.[8,9] Similarly, although application of trigger tools does not depend on self-reporting and can capture unreported events, it may yield many false positives requiring laborious chart review verification and is generally conducted retrospectively. Finally, despite the growing list of quality measures that hospitals are required to calculate for regulatory and payment purposes, it remains unclear whether these measures have improved patient safety or reduced adverse event rates in a meaningful way.[10,11]

To support the operational needs of safety leaders, provide data for cross-institutional comparisons, and drive industry-wide improvement, it would be helpful to develop a core set of patient safety triggers and measures. We evaluated options for efficiently and effectively developing consensus on which metrics to choose and identified the World Café method as our tool. The World Café method is a mechanism to facilitate structured discussions.[12,13] Participants are divided into groups, and each group is assigned a different topic for the purpose of knowledge sharing, consensus building, and/or decision making.

Using the World Café method to derive expert consensus, we sought to generate a contemporary list of triggers and measures of patient harm that could be used 1) for chart review to determine the current incidence of inpatient/outpatient adverse events and 2) to validate electronic tools that monitor for adverse events in real time.[14,15] Beginning with a list of commonly used triggers and National Quality Forum (NQF) endorsed measures, we invited experts from a diverse sample of hospitals to score each on clinical importance, ease of retrospective chart extraction, and suitability for automated electronic chart abstraction.

METHODS

The World Café event was conducted in March 2018 as part of a larger multisite study, which aims to determine the incidence of

From the *Division of Nephrology, Massachusetts General Hospital; †Edward P. Lawrence Center for Quality and Safety, Massachusetts General Hospital; ‡Harvard Medical School; §Division of Pulmonary and Critical Care, Massachusetts General Hospital; ‖Department of Quality and Safety and ¶Division of General Internal Medicine, Brigham and Women's Hospital, Boston; **Division of General Internal Medicine, Mt Auburn Hospital, Cambridge; ††Risk Management Foundation of the Harvard Medical Institutions (CRICO); ‡‡Clinical and Quality Analysis, Mass General Brigham; §§Department of Surgery, Massachusetts General Hospital; ‖‖Harvard T. H. Chan School of Public Health; ¶¶Division of Internal Medicine, Massachusetts General Hospital; and ***Department of Health Care Policy, Harvard Medical School, Boston, MA.

Correspondence: Merranda Logan, MD, MPH, Massachusetts General Hospital, Bulfinch 270c 55 Fruit St, Boston, MA 02114 (e-mail: mslogan@partners.org).

This work was funded by a grant from CRICO (Boston, MA). Dr. Bates consults for EarlySense, which makes patient safety monitoring systems. He receives cash compensation from CDI (Negev), Ltd, which is a not-for-profit incubator for health IT startups. He receives equity from Valera Health, which makes software to help patients with chronic diseases. He receives equity from Clew, which makes software to support clinical decision making in intensive care. He receives equity from MDClone, which takes clinical data and produces deidentified versions of it. He will be receiving research funding from IBM Watson Health. The other authors report no conflicts of interest.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.journalpatientsafety.com).

TABLE 1. Total Number of Triggers and Measures Reviewed by Clinical Domain

Clinical Domains	No. Triggers	No. Measures
1. Ambulatory	39	37
2. Care transitions	5	66
3. Critical care and DVT/PE	8	27
4. Diagnostic/general inpatient	8	77
5. Infection control	3	19
6. Medication/allergies	40	16
7. Nursing-sensitive indicators	3	12
8. Perinatal/maternal	9	46
9. Regulatory/compliance	0	29
10. Surgical/registries	19	62
Total	134	391

DVT, deep vein thrombosis; PE, pulmonary embolism.

inpatient and outpatient adverse events and to develop operational approaches that facilitate the efficient, accurate, and timely measurement of harm. The study team includes 15 investigators from 7 institutions and 6 administrative personnel.

A modified version of the World Café method[12,13] was used to conduct focused discussions on current safety monitoring metrics in 10 clinical domains: ambulatory, care transitions, critical care (including deep venous thrombosis and pulmonary embolus), diagnostic/general inpatient, infection control, medication/allergies, nursing sensitive indicators, perinatal/maternal, regulatory/compliance, surgical/registries. Participants, and alternates in the event of scheduling conflicts, were nominated for their clinical domain expertise by quality/safety leaders at their institutions and were required to have a clinical background (physician, nurse, pharmacist, infection control specialist, etc).

After a group introduction by the leaders of the study team, experts were given 90 minutes to discuss a preidentified list of triggers and measures related to their clinical domain. The list contained the name, a brief description, and the developer or sponsor of each trigger/measure. The experts were asked to rate each measure on clinical importance as a measure of harm, as well as suitability for chart review and electronic extraction, using a 5-point Likert scale (very low, low, medium, high, very high). The experts were instructed to use their clinical background and current work environment to develop ratings. A final rating for each measure was achieved by table consensus. The experts remained at their assigned table for the entirety of the event. Each table was led by an investigator from the study team or designee who received a 1-hour training 1 week before the event. A scribe was assigned to each table to record notes in real time about the group's evaluation of each measure. These were later typed into a spreadsheet by study team personnel. Audio recording was not possible because of noise constraints in the room. Two moderators who were experts in all domains of safety measurement circulated among the tables. At the end of the 90-minute block, experts were asked to share highlights of their conversations, including recurrent themes, with the larger group for 30 minutes.

The preidentified list of triggers/measures was developed by the study team. A total of 134 triggers and 391 measures were included in the master list (range, 15–81 per table; Table 1). All triggers from the Institute for Healthcare Improvement (IHI) trigger tool were included for expert review.[14] In addition, all NQF-endorsed measures were initially considered, but the study team excluded those that were not directly related to patient safety, errors, or adverse events. Because table leads had access to the materials ahead of time, other measures were included if requested by them, particularly if the measure was in current use to support operations.

Because of time constraints at the World Café event, some tables were not able to finish rating all the measures assigned to them, in which cases they were subsequently discussed and rated

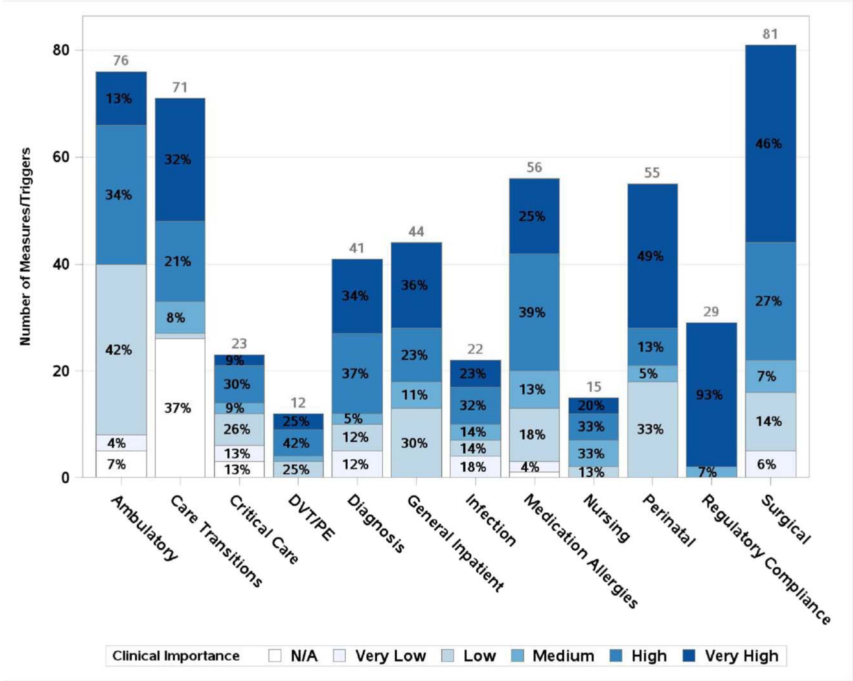

FIGURE 1. Number of measures and triggers by clinical domain and clinical importance. *y* Axis shows count of measures/triggers. Total number of measures for each clinical area is printed above its corresponding bar. Percent of measures with each rating within a clinical area is printed in each segment of the bar. N/A refers to measures deemed repetitive or undesirable that were not scored.

by members of the study team. Table leads from the corresponding clinical domain were then consulted for their opinion on those measures.

RESULTS

We achieved a 72% acceptance rate for our first round of invitations. Seventy-one experts from 9 diverse institutions attended with table size, ranging from 5 to 9 people. Every table had at least 3 institutions represented. Half of the attendees were physicians, 23% were registered nurses, and 7% were pharmacists. The remaining participants were health care professionals of varied backgrounds. Women comprised 59% of the attendees.

Of 525 total triggers and measures, 67% of 391 measures were deemed to have high or very high clinical importance along with 46% of 134 triggers. For those triggers and measures with high or very high clinical importance, 218 overall were deemed to be highly amenable to chart review and 198 for electronic extraction, with an overlap of 192 items suitable for both. Figure 1 shows the assessment of clinical importance by clinical domain, with darker shading indicating higher clinical importance. Figure 2 includes only measures/triggers of high or very high clinical importance and shows the assessment of chart review suitability, with darker shading indicating higher suitability. Figure 3 demonstrates the assessment of measures/triggers with high clinical importance for electronic extraction, with darker shading indicating higher suitability. The surgical table deemed the greatest number of measures clinically important; they also assessed 63% to be very highly amenable to chart review. The regulatory/compliance table deemed the highest percentage of measures, 93%, to be clinically important, whereas the infection table deemed no trigger or measure of high clinical importance to be highly amenable to chart review. The general inpatient table deemed the greatest number of clinically important measures/triggers to be highly suitable for electronic extraction (88%).

Results by table were assembled into individual heat maps, grouping measures by level of clinical importance followed by suitability for chart review and electronic extraction. Figure 4 is an example of the Nursing Sensitive Indicators Table. A scored list of all measures and triggers reviewed, organized by clinical domain, is included in Appendix A, http://links.lww.com/JPS/A336.

DISCUSSION

We evaluated many triggers and metrics, including triggers from IHI and NQF-endorsed measures relating to patient safety. We learned from our World Café process that approximately 6 in 10 were felt to be clinically relevant. There were 322 felt to be of high or very high clinical importance; among those, 218 were suitable for manual chart review and 198 for electronic extraction. Furthermore, of those that are relevant, not all are suitable for manual chart review and/or electronic extraction.

Patient safety leaders depend on a variety of signals as indicators of potential harm. By developing a subset of measures and triggers that can be easily or even automatically extracted, this process has potential for real-time operational monitoring without unrealistic administrative burden. In current practice, manual chart review is the criterion standard method for identifying adverse events, but this approach is time-consuming and expensive.[15–18] Voluntary reporting systems are another approach but capture only a small fraction of all the adverse events that occur.

Trigger tools are an attractive option that do not depend on self-reporting and thus have the potential for capturing unreported events. However, though effective, trigger tools are laborious and traditionally are only applied to a limited sample of retrospective patient records.[14,17,19] The cost of operationalizing a trigger tool-based approach can be substantial, as trained professionals must manually review flagged patient records to confirm or reject whether an adverse event has actually occurred. For all these reasons, we believe that use of the trigger tool alone is an impractical

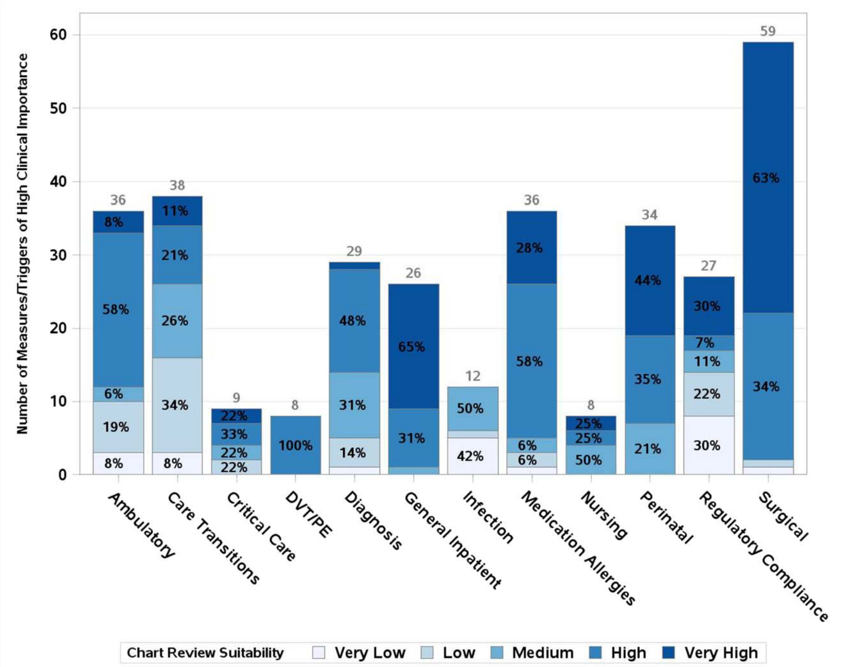

FIGURE 2. Chart review suitability of metrics of high clinical importance by clinical domain. *y* Axis shows count of measures/triggers deemed of high clinical importance. Total number of measures of high clinical importance for each clinical area is printed above its corresponding bar. Percent of measures of high clinical importance with each rating within a clinical area is printed in each segment of the bar.

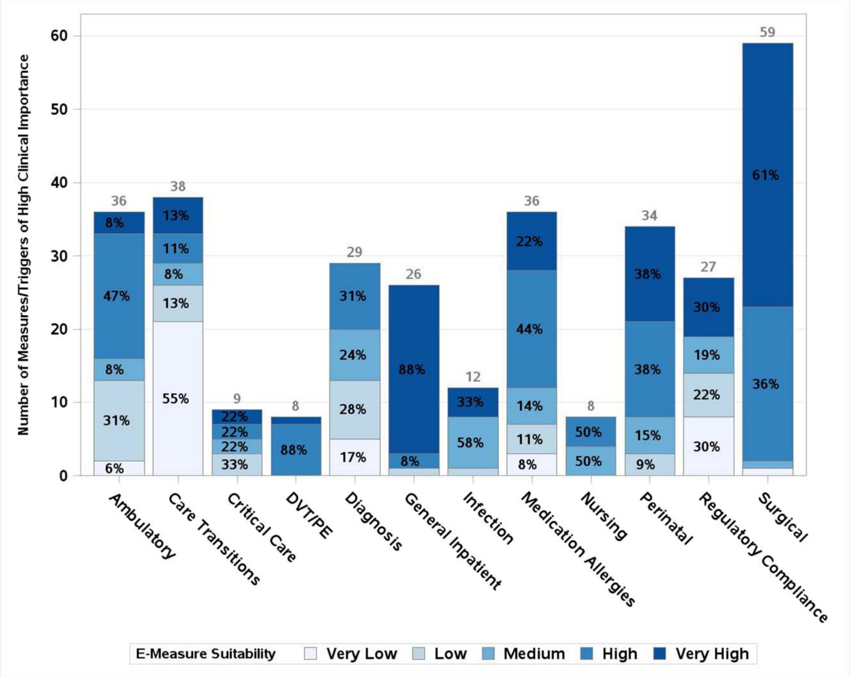

FIGURE 3. Electronic extraction suitability of metrics of high clinical importance by clinical area. *y* Axis shows count of measures/triggers deemed of high clinical importance. Total number of measures of high clinical importance for each clinical area is printed above its corresponding bar. Percent of measures of high clinical importance with each rating within a clinical area is printed in each segment of the bar.

approach to screening entire populations and is best suited for research purposes. A focused list of high-yield triggers is reasonable for routine monitoring of patient safety, but it must be validated.

Rate-based harm measures address some of the shortcomings of trigger tools and can potentially serve as a real-time method for identifying adverse events. They can potentially be applied to an entire population using fewer resources and can theoretically capture some harm events without the need for confirmation by clinician chart review. However, rate-based quality measures must be clinically validated if based on automated extraction from

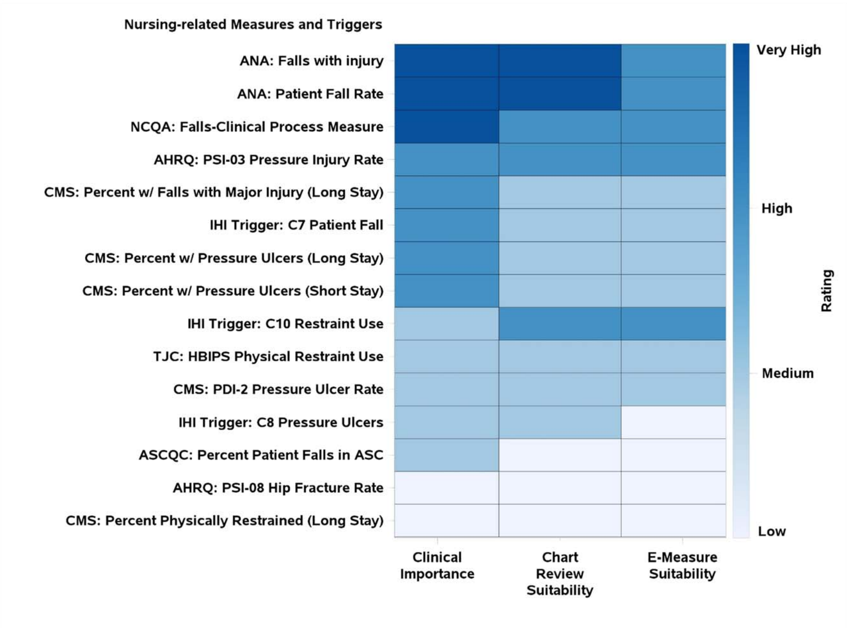

FIGURE 4. World Café measures heatmap (in this example, for nursing measures). No nursing measures were given the rating of "very low." ANA, American Nursing Association; AHRQ, Agency for Healthcare Research and Quality; ASC, Ambulatory Surgery Center; ASCQC, ASC Quality Collaboration; CMS, Centers for Medicare and Medicaid Services; HBIPS, Hospital-Based Inpatient Psychiatric Services; NCQA, National Committee for Quality Assurance; PDI, Pediatric Quality Indicator; PSI, Patient Safety Indicators; TJC, The Joint Commission. Long stay, >100 days; short stay, ≤100 days.

billing data (which have their own intrinsic disadvantages), and this can be as time-consuming and costly as validation of trigger tools. For some clinical areas, national or regional clinical registries produce unadjusted and adjusted rates of important adverse outcomes, but these approaches are also expensive and laborious.

Because of the cost and clinician effort required to collect and validate adverse events using these traditional approaches, electronic clinical quality measures that capture data from the electronic health record (EHR) are emerging and may be used to identify adverse events in real time.[20,21] However, building and validating electronic clinical quality measures can also take considerable time and effort. Furthermore, as evidenced by the results of our World Café, some measures related to patient safety are not feasible for electronic extraction, particularly measures where the numerator and denominator attributes are not captured in a coded EHR field or documented in a structured, standardized manner.

Thus, there is no one perfect method for identifying all adverse events and the industry is currently using mixed methodology to measure patient harm. In addition, the extent of condition-specific measures of harm and measures of harm across the continuum of care varies widely. Some clinical areas, such as surgery, have developed more detailed measures in the context of professional registries, for example, the National Surgical Quality Improvement Program and the Society of Thoracic Surgeon's National Database.[22–24] Our experts deemed 73% of surgical metrics reviewed to be of high or very high clinical importance. Of all the clinical domains reviewed, the regulatory/compliance metrics, comprised primarily of serious reportable events as defined by the NQF (Appendix A, http://links.lww.com/JPS/A336), performed best, with 100% of metrics rated as high or very high clinical importance. This perhaps speaks to the rigorous processes used by the NQF in developing expert consensus. Other areas of clinical care, such as ambulatory care, have fewer validated measures of patient harm and will need more development, perhaps via the emergence of professional registries or consensus panels.[25,26] In contrast to the surgical and regulatory domains, our experts deemed only 47% of the ambulatory metrics reviewed to be of high or very high clinical importance. Combinations of all these approaches should be investigated with the goal of identifying the highest yield measures and screening approaches in every clinical domain.

Our World Café results suggest that experts think that many quality measures are not important. The focus of adverse event detection should be on measures that are clinically relevant, capable of being captured accurately and consistently with the lowest resource use, publicly reported, or tied to reimbursement. Such measures will capture the greatest attention of both clinicians and senior hospital leaders and will hopefully be less likely to lead to unintended negative impacts on behavior and outcomes.[27–29] In a 2014 study by Lindenauer et al[30] of senior leaders from 280 US hospitals, roughly half of the leaders surveyed did not believe that publicly reported measures accurately portrayed the quality of care or were useful to infer hospital quality. Furthermore, senior hospital leaders expressed concern that the focus on publicly reported quality measures might lead to neglect of other clinically important matters.[30] Balancing clinical importance with ease of calculation is especially important for publicly reported measures.

Our study must be interpreted in the context of potential limitations. Although we performed a comprehensive literature review, we primarily focused on the IHI trigger tool and measures endorsed by NQF; other comparable systems may be available. Our ratings for clinical importance should be interpreted as the consensus opinion of our diverse panel of experts. Our ratings for ease of manual and electronic extraction should be interpreted in the context of our documentation culture and the design of our EHRs. Finally, a representative sample of Massachusetts hospitals were included in the World Café, but not all EHRs were represented.

CONCLUSIONS

The dramatic expansion of health care quality and safety measures and the variation in the use of these measures across sites have created both opportunities and challenges. There are far too many valid measures, many of which have overlapping focus and are redundant. Other desirable measures are far too costly and time-consuming to be collected routinely, longitudinally, and in near real-time. The goal of our World Café was to identify those measures that seemed most relevant to clinical and health policy experts, which had the greatest potential for automated extraction from the EHR or administrative sources and which would require the least manual collection or validation. Our hope is to use this measure set to explore the prevalence of patient harm in today's health care delivery system and use this as a basis on which to suggest a measure set for ongoing surveillance and improvement for the future.

Lessons learned from this exercise will be used for the next phase of our project. The measures of high or very high clinical importance that we identified will be collected from a sample of inpatient and outpatient records from hospitals of varying size and teaching intensity. The results of this exercise will further refine our selection of optimal measures for monitoring adverse events. Finally, we will investigate and validate automated approaches to extracting the information required for these measures using computer-based and machine-learning approaches.

REFERENCES

1. Kohn LT, Corrigan JM, Donaldson MS. *To Err Is Human*. Washington, DC: The National Academies Press; 2000.
2. Institute of Medicine & Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century.* The National Academies Press; 2000;2001.
3. Landrigan CP, Parry GJ, Bones CB, et al. Temporal trends in rates of patient harm resulting from medical care. *N Engl J Med*. 2010;363:2124–2134.
4. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. *BMJ.* 2016;353:i2139.
5. Leape LL, Brennan TA, Laird N, et al. The nature of adverse events in hospitalized patients. *N Engl J Med*. 1991;324:377–384.
6. Brennan TA, Leape LL, Laird NM, et al. Incidence of adverse events and negligence in hospitalized patients. *N Engl J Med*. 1991;324:370–376.
7. Thomas EJ, Petersen LA. Measuring errors and adverse events in health care. *J Gen Intern Med*. 2003;18:61–67.
8. Cullen DJ, Bates DW, Small SD, et al. The incident reporting system does not detect adverse drug events: a problem for quality improvement. *Jt Comm J Qual Improv.* 1995;21:541–548.
9. Rothschild JM, Landrigan CP, Cronin JW, et al. The critical care safety study: the incidence and nature of adverse events and serious medical errors in intensive care. *Crit Care Med*. 2005;33:1694–1700.
10. Jha AK, Joynt KE, Orav EJ, et al. The long-term effect of premier pay for performance on patient outcomes. *N Engl J Med*. 2012;366:1606–1615.
11. Ryan AM, Krinsky S, Maurer KA, et al. Changes in hospital quality associated with hospital value-based purchasing. *N Engl J Med*. 2017;376: 2358–2366.
12. Brown J, Isaacs D. *The World Café : Shaping Our Futures through Conversations That Matter.* Oakland: Berrett-Koehler Publishers; 2005.
13. World Cafe Method: The World Cafe. Available at: http://www.theworldcafe.com/key-concepts-resources/world-cafe-method/. Accessed October 9, 2018.

14. Griffin F, Resar R. IHI Global Trigger Tool for Measuring Adverse Events (Second Edition). IHI Innovation Ser White Paper. Cambridge, MA; Institute for Healthcare Improvement. 2009.

15. De Wet C, Bowie P. Screening electronic patient records to detect preventable harm: a trigger tool for primary care. *Qual Prim Care*. 2011;19: 115–125.

16. James JT. A new, evidence-based estimate of patient harms associated with hospital care. *J Patient Saf*. 2013;9:122–128.

17. Unbeck M, Lindemalm S, Nydert P, et al. Validation of triggers and development of a pediatric trigger tool to identify adverse events. *BMC Health Serv Res*. 2014;14:655.

18. Eggleton KS, Dovey SM. Using triggers in primary care patient records to flag increased adverse event risk and measure patient safety at clinic level. *N Z Med J*. 2014;127:45–52.

19. Resar RK, Rozich JD, Classen D. Methodology and rationale for the measurement of harm with trigger tools. *BMJ Qual Saf*. 2003;12: 39ii–3945ii.

20. Tang PC, Ralston M, Arrigotti MF, et al. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Inform Assoc*. 2007;14:10–15.

21. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev*. 2010;67:503–527.

22. ACS National Surgical Quality Improvement Program. Available at: https://www.facs.org/quality-programs/acs-nsqip. Accessed March 19, 2019.

23. Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Ann Surg*. 1998;228:491–507.

24. Jacobs JP, Shahian DM, D'Agostino RS, et al. The Society of Thoracic Surgeons National Database 2017 Annual Report. *Ann Thorac Surg*. 2017; 104:1774–1781.

25. Gandhi TK, Lee TH. Patient safety beyond the hospital. *N Engl J Med*. 2010;363:1001–1003.

26. Gandhi TK, Seger AC, Overhage JM, et al. Outpatient adverse drug events identified by screening electronic health records. *J Patient Saf*. 2010;6: 91–96.

27. Baker DW, Qaseem A, American College of Physicians' Performance Measurement Committee. Evidence-based performance measures: preventing unintended consequences of quality measurement. *Ann Intern Med*. 2011;155:638–640.

28. Wachter RM, Flanders SA, Fee C, et al. Public reporting of antibiotic timing in patients with pneumonia: lessons from a flawed performance measure. *Ann Intern Med*. 2008;149:29–32.

29. Wadhera RK, Joynt Maddox KE, Wasfy JH, et al. Association of the hospital readmissions reduction program with mortality among Medicare beneficiaries hospitalized for heart failure, acute myocardial infarction, and pneumonia. *JAMA*. 2018;320:2542–2552.

30. Lindenauer PK, Lagu T, Ross JS, et al. Attitudes of hospital leaders toward publicly reported measures of health care quality. *JAMA Intern Med*. 2014; 174:1904–1911.