The Elusive and Illusive Quest for Diagnostic Safety Metrics

Gordon D. Schiff, MD[1,2] and Elise L. Ruan, MD, MPH[3]

[1]Harvard Medical School Center for Primary Care, Boston, MA, USA; [2]Brigham and Womens Hospital Center for Patient Safety Research and Practice, Boston, MA, USA; [3]Tufts University School of Medicine, Boston, MA, USA.

J Gen Intern Med 33(7):983–5
DOI: 10.1007/s11606-018-4454-2

Not everything that counts can be counted, and not everything that can be counted counts.

Variously attributed to Albert Einstein, William Bruce Cameron, Lord Platt, and others[1]

Can't improve what you can't measure? Nonsense. Over the decades my relationship with my wife has continuously improved. But I've never administered a survey to her, nor tracked metrics of our relationship. Not only was this not needed for improvement, but likely would have been detrimental and disrespectful.

Don Berwick speaking at Institute for Healthcare Improvement Forum[2]

Diagnosis errors have come out of the periphery of the patient safety movement. With the publication of 2015 National Academy of Medicine Report Improving Diagnosis in Health Care and recent reports suggesting diagnostic errors are the leading types of errors reported by patients and a top reason they file malpractice suits, diagnostic errors are finally gaining the "respect" they warrant.[3,4,5] In the current issue of JGIM, three leading voices in the movement to improve diagnosis propose a framework that they argue will help advance metrics of diagnostic performance within and across health care systems as well as enable researchers and systems to determine the impact of improvement interventions.[6]

The goal of developing and reporting standardized measures related to diagnostic safety has been an elusive one. Despite the urging of multiple organizations and advocates, crafting measures for diagnosis quality has not proven to be simple.[7] But how will we know if we are making progress and how can we hold organizations and clinicians accountable without some objective measures?

To overcome a host of past difficulties in creating such metrics, the authors propose a framework with seven criteria for designing measures of "Undesirable Diagnostic Events" (UDE's). They suggest six diagnoses as logical candidates for places to start. (Olson, Table 1). One clue that this may not be so simple is the fact that in their article, Olsen et al. mention twice that number of diagnoses as examples that would *not* lend themselves to the UDE measurement framework, including herpes zoster, pneumothorax, adult onset Stills, amyloid, Alzheimer's, depression, spinal metastasis, mitochondrial disorders, bacterial overgrowth, adrenal insufficiency, and certain psychiatric conditions.[6] Perhaps just by sheer coincidence, one of us (GS) has personally had two of these (zoster, pneumothorax) *misdiagnosed* by skilled physicians (in addition to initially self-misdiagnosing). Thus this list is revealing not only because it suggests several personally experienced diagnostic failures would be outside the purview of the UDE framework, but we suspect that applying their criteria strictly for the type "never-event" UDE's they advocate would exclude most of the diagnostic errors and problems in the diagnostic process that are occurring in healthcare today.

Let us examine just one of the diagnoses they suggest *would* be a good candidate, tuberculosis. TB is indeed important, being highly prevalent worldwide, as well as an important diagnosis not to miss or delay. Consider the consequences of overlooking a hospitalized patient with active pulmonary TB, both in terms of exposure of other patients and health workers, as well as failure or delay in treating a seriously ill patient with lifesaving medications. So how should we go about designing the proposed UDE performance metric? Should we only measure pulmonary TB? If we did, we would be excluding TB meningitis, miliary TB, and renal and spinal TB—all serious diagnoses not to miss. What about active vs. latent TB which is often difficult to differentiate. What about the false negative and false positive rates of various TB sputum, skin, and blood tests—how should we factor this in when evaluating care quality? The authors mention the finding of TB on autopsy would be the gold standard basis for this metric. However, autopsies are rarely performed in the USA, and subject to serious selection bias that would markedly limit the utility of finding missed TB as an accurate and fair diagnostic performance measure. Finally, one would hope we do not have to wait for a patient to die to uncover diagnostic improvement opportunities.

The purpose of raising these questions is not to nitpick or deny that smarter minds can overcome some of these technical challenges in crafting a TB, or other diagnostic performance measures. Rather it is to raise more fundamental questions that those of us in the diagnostic safety movement as well as

Published online April 30, 2018

clinicians and patients need consider. How will these metrics help us move forward, and importantly how will they positively engage clinicians to achieve the goal of more reliable and timely diagnosis?

OUTSIDE IN VS. INSIDE OUT MEASUREMENT

Quality guru W. Edwards Deming is reported to have said "when I see workers measuring themselves, I see quality." In this profound statement, he was both downplaying the value of external measurement as well as extolling the importance of motivated, empowered workers trained in self-measurement skills (e.g., using statistical process control (SPC) charts to differentiate "special cause" (special circumstances, unexpected outlier defects) from "common cause" variation (random variation that is part of the system) taking the initiative to examine and improve their own quality. There is an even deeper significance to this concept of ensuring quality by creating a culture where front line staff, rather than external inspection or metrics, are the key to safe diagnosis. To explain, consider an analogy to the modern day approaches to ensuring medication quality.

For decades, US Pharmacopeia (USP), the official certifier of chemical quality for drug products marketed in the USA, relied primarily on sophisticated laboratory methods for inspecting drug product samples produced and submitted by each manufacturer. Using established laboratory techniques such as chromatography, USP scientists would check the purity and strength of the ingredients of these samples against reference standards to ensure they conformed to the strict standards that had been established for that drug entity. However, increasingly this inspection approach to medication quality has been displaced by a very different approach—continuous process verification, whereby continuous assurance is available to detect any unplanned departures and allow manufacturers to identify and adjust for them, thus helping prevent product failures. USP standards now provide precise formulas and preparation guidelines, along with pure reference samples for testing, so that drugs can be made consistently, every time. A similar approach is the basis for ISO (International Organization for Standardization) good manufacturing standards which "provide requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose." In short, quality, safety, and efficacy are designed/built into the product, not "inspected-in" after the fact.

Perhaps we need to apply these state-of-the-art manufacturing constructs to improving diagnosis. What would be the steps to specify "well-made" diagnoses? For example, we should certainly want to ensure every test result was returned, acknowledged, and acted on as well as communicated to the patient by the ordering clinician. Although this processes standard would appear to be a relatively straightforward one (at least compared to other more complex aspects of diagnosis), we now know that both specifying and ensuring such a reliable closed loop system is not a trivial feat.

Under what conditions should diagnoses be produced (really *co-produced)* with patients? Given what we know about the importance of safety culture in general, could we specify (and even measure) organizational expectations, processes, and conditions for assuring quality diagnoses and learning from errors? (Text Box 1) Compare an organization that relied on publicly reporting of its "never event" numbers for the six proposed UDE diagnoses vs. one that was hard-wired with meaningful process and learning culture and safeguards embodied in our diagnosis culture framework? In which organization/system would you feel safer and more confident that the safeguards were in place for reliable timely diagnosis? In which system would you rather be a patient? And in which would you rather work?[8]

Text Box 1. Culture of diagnostic safety and improvement

- Replacing blame and fear with learning and improvement, so no one is afraid to ask questions, question a diagnosis, or transparently share when things go wrong
- Commitment to improving diagnosis, learning from delays and diagnostic process errors
  - Organizational recognition that misdiagnosis is #1 cause of patient-reported errors
  - Comprehensive reporting, appreciative investigation of adverse events
  - Relentless curiosity/worry/conferencing: what we might be missing, what can go wrong in system
  - Attention to details of diagnostic process and what can go wrong, awareness of limitations of tests
- Recognition that uncertainty is inherent in diagnoses, tests, illness presentation, and evolution; anticipation of common pitfalls
  - Situational awareness of local, disease-specific, literature-reported vulnerabilities and pitfalls
  - Hard-wired, proactive, reliable follow-up safety nets, and feedback systems to detect and protect
  - Conservative approaches to testing and imaging, enabled by shared decision-making and reliable follow-up
- Respect for human limitations and need for cognitive, process support
  - Decreased reliance on human memory, minimizing negative effects of stress, fatigue, fear, appreciating risks of multitasking
  - Redesign of EMR and communication systems to support cognition, collaborative diagnosis, follow-up
- Enhanced role for patients in co-producing diagnosis
  - Working collaboratively to formulate history, diagnosis, monitor course, and raise and research questions

We now know that so-called quality reporting (particularly those more market-oriented approaches to encourage patients to shop around for quality or financially incentivize (or punish) institutions based on their performance) is vulnerable to a myriad of issues including problems with measurement, case mix adjustment, incentives to game measures to make performance look better than it is, neglect of areas (in this case different diagnoses) not covered, clinician cynicism and skepticism with "box-ticking" and ill-informed second guessing, and time and resources required to collect (often manually) data for pubic reporting of dubious proven value that present an incomplete picture of clinicians' diagnostic work.[9,10,11] These two approaches-metrics vs. culture– of course are not mutually exclusive requiring us to make an either/or choice.

And in many ways, narrower outcome metrics and culture could work together in a complementary fashion. However, before we go down this "metrics" road, we need to critically weigh what such measurements will and will not bring to improving diagnosis.

Perhaps focusing more closely on collaborative learning from the stories and details of actual cases of diagnostic error can be a more powerful lever for accountably and improvement than bar graphs or pie charts.[12, 13] The success of the #MeToo movement in exposing and limiting sexual misconduct demonstrates the power of impact over metric. Qualitatively understanding the plethora of diagnostic errors locally and across institutions can help us build the situational awareness and safety nets we need for better diagnostic conduct.

Acknowledgements: *The authors acknowledges support for research in diagnostic error and improvement from CRICO (Harvard affiliated organizations malpractice insurer) and the Gordon and Betty Moore Foundation.*

Corresponding Author: *Gordon D. Schiff, MD; Harvard Medical School Center for Primary Care, Boston, MA, USA (e-mail: gschiff@bwh.harvard.edu).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors have no financial conflicts.*

REFERENCES

1. Investigator Q. Not Everything That Counts Can Be Counted. 2010; https://quoteinvestigator.com/2010/05/26/everything-counts-einstein/#note-455-9. Accessed 1/15/20118, 2018.
2. **Berwick DM**. *Escape fire: designs for the future of health care*. John Wiley & Sons; 2010.
3. **Balogh E, Miller BT, Ball J.** *Improving diagnosis in health care*. National Academies Press; 2015.
4. **Schiff GD, Puopolo AL, Huben-Kearney A, et al.** Primary care closed claims experience of Massachusetts malpractice insurers. *JAMA internal medicine*. 2013;173(22):2063–2068.
5. The Public's Views on Medical Error in Massachusetts. https://cdn1.sph.harvard.edu/wp-content/uploads/sites/94/2014/12/MA-Patient-Safety-Report-HORP.pdf.
6. **Olson A, Graber, Mark, Singh, H.** Tracking Progress in Improving Diagnosis: A Framework for Defining Undesirable Diagnostic Events *JGIM* 2018. SPI 4340.
7. National Quality Forum. *Improving Diagnostic Quality and Safety Draft Measurement Framework*. Washington DC: National Quality Forum 2017
8. **Berwick DM.** Continuous improvement as an ideal in health care. In: Mass Medical Soc; 1989.
9. **Greener I, Harrington B, Hunter D, Mannion R, Powell M.** A realistic review of clinico-managerial relationships in the NHS: 1991-2010. *National Institute for Health Research, Service Delivery & Organisation programme*. 2011.
10. **Himmelstein D, Woolhandler S.** Quality improvement:"Become good at cheating and you never need to become good at anything else.". *Health Affairs Blog*. 2015.
11. **Casalino LP, Nicholson S, Gans DN, et al.** What does it cost physician practices to interact with health insurance plans? *Health Affairs*. 2009;28(4):w533-w543.
12. **Hoff, Timothy J.** *Next in Line: Lowered Care Expectations in the Age of Retail-and Value-based Health*. Oxford University Press, 2017.
13. **Berwick DM.** The stories beneath. *Medical care*. 2007;45(12):1123–1125.