A Follow-Up Report Card on Computer-Assisted Diagnosis—the Grade: C+

Craig A. Umscheid, MD, MSCE, FACP^{1,2,3,4} and C. William Hanson, MD, III^{2,5,6}

¹Medicine and Epidemiology, Center for Evidence-based Practice, University of Pennsylvania, Philadelphia, PA, USA; ²Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA; ³Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA, USA; ⁴Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, USA; ⁵Department of Anesthesiology and Critical Care, University of Pennsylvania, Philadelphia, PA, USA; ⁶Office of the Chief Medical Information Officer, University of Pennsylvania, Philadelphia, PA, USA.

J Gen Intern Med 27(2):142–4

DOI: 10.1007/s11606-011-1944-x

© Society of General Internal Medicine 2011

Diagnostic errors result in an estimated 40,000 to 80,000 US hospital deaths annually¹; they are the most common type of medical error cited in malpractice claims and result in the largest payouts²; they lead to more patient readmissions than other types of errors³; and autopsy studies confirm the ubiquity of missed diagnoses.⁴ Although not all diagnostic errors result in harm, many do.⁵ Typically these mistakes represent either cognitive errors, such as those resulting from a lack of knowledge, or communication errors, such as those resulting from poor handoffs and follow-up.^{6,7} Differential diagnosis (DDX) generators have the potential to assist in the former, while other types of clinical decision support can address the latter.

In this issue of JGIM, Bond and colleagues examine the features and performance of four DDX generators.⁸ They identified these tools through a systematic search, compared each tool against a set of consensus criteria, and tested each tool using diagnostic conundrums from the New England Journal of Medicine (NEJM) and the Medical Knowledge Self Assessment Program (MKSAP) of the American College of Physicians. The best tools demonstrated sensitivities as high as 65% and a number of features which advanced their usability, including direct links to evidence resources, the ability to input multiple variables including negative findings, the potential to be integrated with electronic health records (EHRs) to obviate the need for manual data entry, and mobile access.

Unfortunately, the performance results were not much better than programs examined in a similar study published in the NEJM nearly 20 years ago by Berner and colleagues.⁹ In that study, four DDX generators were evaluated, two of which are no longer on the market, and only one of which met the inclusion criteria of the current study. Berner and colleagues concluded that on average the tools had a relatively mediocre sensitivity of approximately 50–70% for the correct diagnosis, and a companion editorial by Kassirer gave the tools a grade of C.¹⁰

The current study by Bond and colleagues has many strengths. First, their search for DDX generators was systematic and comprehensive, including searches of the internet and medical literature databases, as well as expert consultation. The appendix includes a list of programs which were identified

but did not meet inclusion criteria. The only major tool type that was not considered for inclusion in their study was the internet search engine (such as the generic Google or the medically targeted WebMD). Evidence suggests that these search tools can lead to the correct diagnosis in over 50% of cases, similar to the sensitivities of the DDX generators in this study.¹¹ Thus, head-to-head comparisons of DDX generators with internet search engines will be critical to demonstrating the incremental value of the proprietary, specialized software. Of note, at least one of the tools examined in Bond's study offered structured Google searching of pre-selected medical websites.

A second strength of the study is the consensus-based evaluation template developed by the authors. The attributes had real-world face validity, and included critical measures like: 1) the potential for EHR integration, 2) Health Level 7 (HL7) interoperability, 3) options for mobile access, 4) the types of data inputs allowed (for example, labs, drugs, images, geographic location, and negative findings), 5) use of natural language for queries, 6) source of knowledge content, 7) links to Pubmed and other evidence-based content, 8) frequency of content updates, and 9) the potential for usage tracking. The only downside of this list of attributes is that the authors assessed some of the attributes using manufacturer's claims rather than direct testing of the DDX generators. This is a particularly important limitation for an attribute like EHR integration, which would require testing in one's local setting to verify manufacturers' claims.

Another strength of the study lies in the methods used to test the DDX generators. Diagnostic cases were appropriately varied and challenging, and included pheochromocytoma, stress cardiomyopathy (Takotsubo), amyloidosis, Goodpastures syndrome, and copper deficiency, among others (see Appendix 2 of Bond's study for the complete list). The selection of cases is important as physicians are unlikely to resort to a DDX generator for 'routine' cases such as community acquired pneumonia. The authors also ensured that the researchers choosing the findings to enter from each case and entering the findings into the DDX generator were blinded with regards to the case diagnosis. Moreover, the 5-point performance-scoring system included gradations of "correctness" or "helpfulness", from a 5 (in which the correct diagnosis was included in the top 20 diagnoses generated) to a 0 (in which no diagnosis in the top 20 was relevant to the target diagnosis).

Despite these strengths, the study by Bond and colleagues suffers from a critical limitation: there is no assessment of the

impact of these products on true clinical outcomes.¹² Thus, we are left to wonder whether use of the examined tools would decrease diagnostic error, or more importantly misdiagnosis-related harm.¹ In addition, we are left to wonder if use of DDX generators might increase diagnostic testing, with the potential for associated complications and costs. Another potential hazard is false-confidence: providers might become more attached to their initial diagnosis merely because it appeared on a generated list of diagnoses. As a result, they might be less inclined to get a consult when necessary. There is also no assessment of how DDX generators might affect workflow, length of stay, readmissions, patient satisfaction, and costs.

The impact of these tools might also be affected by the patient care setting in which they are applied. For example, one could imagine that academic medical centers, with their mix of diagnostic dilemmas, patient complexity, and young trainees, might benefit more from DDX generators than community hospitals with more “bread and butter” presentations. On the flip side, maybe such products would be dangerous in the hands of young trainees for the reasons noted above. Providers in cognitive fields like internal medicine might benefit more than surgical colleagues. Cognitive fields suffer more from diagnostic error than surgical fields, so a differential impact of these tools by specialty might well be expected.³

Another limitation of the study by Bond is the lack of usability testing in the real-world. The question of whether DDX generators would actually be used, and if so when and by whom, is of paramount importance. An assessment of the ease of use of these programs and user satisfaction with the interface would be a first step to understanding program usability. But more importantly, studies need to assess whether those who need assistance from these programs would actually use them. Studies suggest that provider overconfidence^{13,14} and premature closure¹⁵ are both associated with diagnostic error. So it is unclear if DDX generators can reduce misdiagnosis-related harm if we simply rely on providers to use these programs when they deem necessary, or if more active interfaces or prompts would be necessary to have an impact on clinical outcomes. At the heart of all of these questions is the provider's perception of the incremental value of such tools.

A 2006 study by Ramnarayan and colleagues represents a good first step to examining the impact of DDX generators in the real world.¹⁶ The study examined the use by pediatric housestaff of a DDX generator over a 5-month period in four outpatient pediatric clinics. As part of the study, users were compelled to enter the differential diagnoses, tests and treatments they were considering before engaging the tool. They were again asked to record their management plans following use of the DDX generator. The study showed that housestaff “attempted” to use the tool on 595 occasions (about 9% of all patient encounters), but only “examined” the diagnostic advice from the tool in 177 episodes (or 30% of the encounters for which use of the tool was “attempted”). For the 104 episodes of care for which complete medical records were available, an expert panel found a significant reduction in “unsafe” diagnostic workups (defined in the study as a deviation from a “minimum gold standard”), from a staggering 45% of episodes to a still concerning 33% of episodes. Median time spent on the DDX generator was about one and a half minutes. User satisfaction with the tool was low, which may have been due to the onerous requirements of the study protocol.

Importantly, none of the DDX generators examined in the current study by Bond address non-cognitive causes of diagnostic error like failures in information transfer. Given the increasing number of patient handoffs in today's medical environment, perhaps electronic solutions targeting this area could reduce misdiagnosis-related harm even more than the ideal DDX generator.^{2,17–19} Potential solutions to address these system errors include EHRs with fail-safe communication for test ordering and results tracking, such that important test results don't fall through the cracks. Automated feedback loops including patients as well as primary and specialty physicians may also help to address errors of information transfer.

Despite the nearly 20 years that have passed between the study by Berner and the current study by Bond, little evidence has justified raising the grade of C that Kassirer originally gave to computer-assisted diagnosis. Given the times, with rampant diagnostic testing and exploding medical costs, one might even claim that currently the potential risks of using DDX generators are greater than in the past, particularly since we still lack any evidence of benefit in clinical outcomes.^{20,21} Yet, given some of the potential improvements these programs have made in the areas of EHR integration, mobile access, allowable data inputs, and links to comprehensive evidence-based resources, we feel compelled to raise the initial grade of C to a C+. Call it grade inflation.

Acknowledgments: No potential conflicts to disclose. No specific sources of funding were used.

Corresponding Author: Craig A. Umscheid, MD, MSCE, FACP; Medicine and Epidemiology, Center for Evidence-based Practice, University of Pennsylvania, 3535 Market Street, Mezzanine, Suite 50, Philadelphia, PA 19104, USA (e-mail: craig.umscheid@uphs.upenn.edu).

REFERENCES

1. Newman-Toker DE, Pronovost PJ. Diagnostic errors—the next frontier for patient safety. *JAMA*. 2009;301(10):1060–1062.
2. Singh H, Naik AD, Rao R, Petersen LA. Reducing diagnostic errors through effective communication: harnessing the power of information technology. *J Gen Intern Med*. 2008;23(4):489–494.
3. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. *Arch Intern Med*. 2010;170(12):1015–1021.
4. Shojania KG, Burton EC, McDonald KM, Goldman L. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *JAMA*. 2003;289(21):2849–2856.
5. Pronovost PJ, Colantuoni E. Measuring preventable harm: helping science keep pace with policy. *JAMA*. 2009;301(12):1273–1275.
6. Gandhi TK, Kachalia A, Thomas EJ, Puopolo AL, Yoon C, Brennan TA, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. *Ann Intern Med*. 2006;145(7):488–496.
7. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med*. 2005;165(13):1493–1499.
8. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential Diagnosis Generators: an Evaluation of Currently Available Computer Programs. *J Gen Intern Med*. 2011; doi:10.1007/s11606-011-1804-8.
9. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. *N Engl J Med*. 1994;330(25):1792–1796.
10. Kassirer JP. A report card on computer-assisted diagnosis—the grade: C. *N Engl J Med*. 1994;330(25):1824–1825.

11. **Tang H, Ng JH.** Googling for a diagnosis—use of Google as a diagnostic aid: internet based study. *BMJ*. 2006;333(7579):1143–1145.
12. **Fryback DG, Thornbury JR.** The efficacy of diagnostic imaging. *Med Decis Making*. 1991;11(2):88–94.
13. **Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS, et al.** Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. *J Gen Intern Med*. 2005;20(4):334–339.
14. **Berner ES, Graber ML.** Overconfidence as a cause of diagnostic error in medicine. *Am J Med*. 2008;121(5 Suppl):S2–S23.
15. **Croskerry P.** The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med*. 2003;78(8):775–780.
16. **Ramnarayan P, Winrow A, Coren M, Nanduri V, Buchdahl R, Jacobs B, et al.** Diagnostic omission errors in acute paediatric practice: impact of a reminder system on decision-making. *BMC Med Inform Decis Mak*. 2006;6:37.
17. **Schiff GD, Bates DW.** Can electronic clinical documentation help prevent diagnostic errors? *N Engl J Med*. 2010;362(12):1066–1069.
18. **Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW.** Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. *JAMA*. 2007;297(8):831–841.
19. **Riesenberg LA, Leitzsch J, Massucci JL, Jaeger J, Rosenfeld JC, Patow C, et al.** Residents' and attending physicians' handoffs: a systematic review of the literature. *Acad Med*. 2009;84(12):1775–1787.
20. **Amsterdam EA, Kirk JD, Bluemke DA, Diercks D, Farkouh ME, Garvey JL, et al.** Testing of low-risk patients presenting to the emergency department with chest pain: a scientific statement from the American Heart Association. *Circulation*. 2010;122(17):1756–1776.
21. **Chou R, Gaseem A, Owens DK, Shekelle P.** Diagnostic imaging for low back pain: advice for high-value health care from the American College of Physicians. *Ann Intern Med*. 2011;154(3):181–189.