REVIEWS

# Differential Diagnosis Generators: an Evaluation of Currently Available Computer Programs

*William F. Bond, MD, MS[1,2], Linda M. Schwartz, MDE[2], Kevin R. Weaver, DO[1], Donald Levick, MD, MBA[3,4], Michael Giuliano, MD, MEd, MHPE[5], and Mark L. Graber, MD[6]*

[1]Department of Emergency Medicine, Lehigh Valley Health Network, Allentown, PA, USA; [2]Division of Education, Lehigh Valley Health Network, Allentown, PA, USA; [3]Department of Information Services, Lehigh Valley Health Network, Allentown, PA, USA; [4]Department of Pediatrics, Lehigh Valley Health Network, Allentown, PA, USA; [5]Department of Pediatrics, Hackensack University Medical Center, Hackensack, NJ, USA; [6]Department of Medicine Veterans Administration Medical Center, Northport, NY, USA.

***BACKGROUND:*** Differential diagnosis (DDX) generators are computer programs that generate a DDX based on various clinical data.

***OBJECTIVE:*** We identified evaluation criteria through consensus, applied these criteria to describe the features of DDX generators, and tested performance using cases from the New England Journal of Medicine (NEJM©) and the Medical Knowledge Self Assessment Program (MKSAP©).

***METHODS:*** We first identified evaluation criteria by consensus. Then we performed Google® and Pubmed searches to identify DDX generators. To be included, DDX generators had to do the following: generate a list of potential diagnoses rather than text or article references; rank or indicate critical diagnoses that need to be considered or eliminated; accept at least two signs, symptoms or disease characteristics; provide the ability to compare the clinical presentations of diagnoses; and provide diagnoses in general medicine. The evaluation criteria were then applied to the included DDX generators. Lastly, the performance of the DDX generators was tested with findings from 20 test cases. Each case performance was scored one through five, with a score of five indicating presence of the exact diagnosis. Mean scores and confidence intervals were calculated.

***KEY RESULTS:*** Twenty three programs were initially identified and four met the inclusion criteria. These four programs were evaluated using the consensus criteria, which included the following: input method; mobile access; filtering and refinement; lab values, medications, and geography as diagnostic factors; evidence based medicine (EBM) content; references; and drug information content source. The mean scores (95% Confidence Interval) from performance testing on a five-point scale were Isabel© 3.45 (2.53, 4.37), DxPlain® 3.45 (2.63–4.27), Diagnosis Pro® 2.65 (1.75–3.55) and PEPID™ 1.70 (0.71–2.69). The number of exact matches paralleled the mean score finding.

***CONCLUSIONS:*** Consensus criteria for DDX generator evaluation were developed. Application of these criteria as well as performance testing supports the use of DxPlain® and Isabel© over the other currently available DDX generators.

*KEY WORDS:* differential diagnosis; clinical decision support systems; diagnostic errors; evidence-based medicine; computer-assisted diagnosis.

J Gen Intern Med 27(2):213–9

DOI: 10.1007/s11606-011-1804-8



***Electronic supplementary material*** *The online version of this article (doi:10.1007/s11606-011-1804-8) contains supplementary material, which is available to authorized users.*

*Received July 30, 2011*
*Revised February 28, 2011*
*Accepted June 24, 2011*
*Published online July 26, 2011*

## BACKGROUND

Diagnostic error can lead to inappropriate or absent therapeutic interventions, and thus has substantial human costs for patients. It is one of the most common reasons for malpractice lawsuits and accounts for the largest dollar losses amongst these cases[1,2]. Diagnostic error remains one of the more challenging areas of patient safety because of the hidden nature of cognitive processing and the many factors (affective, patient-related, environmental, and systems-related) that influence medical decision making[3,4]. The challenge to practicing clinicians is to prevent misdiagnosis in real time, and to teach this skill to trainees. Thus, any proactive support system that would help clinicians in teaching or executing the medical diagnostic decision-making process would be welcome[5].

Differential diagnosis (DDX) generators are computer programs that assist the clinician by combining symptoms, findings, and other factors to suggest a list of possible diagnoses for consideration. Computer-assisted differential diagnosis generation has been available since the mid-1980s[6]. One of the most important works evaluating the performance of DDX generators was conducted by Berner et al. in 1994. That landmark study pitted four programs against 105 "diagnostically challenging" cases that were created through a consensus process by experts. At that time the simple presence of the primary case diagnosis within the possible choices of the DDX program list varied in proportion from 0.73 to 0.91 and the proportion of correct diagnoses when the test cases were applied ranged from 0.52 to 0.71. This measure of the correct diagnosis

with test case application is akin to sensitivity. Scores were generated for correct or closely related diagnoses found by the programs, comprehensiveness of the diagnosis list, relevance of the diagnosis list and the presence of useful but previously unconsidered diagnoses. By addressing relevance of the diagnostic list, Berner et al. were touching upon the concept of specificity. However, the programs are designed to generate diagnostic possibilities, and therefore by nature are focused on sensitivity (presence of the diagnosis for the case) rather than specificity (absence of irrelevant diagnoses). The programs were judged to be roughly equivalent in their usefulness. At the time, it was noted that their ability to be useful in practice had yet to be proven[7,8].

A more recent study showed that when presented with the key findings of difficult cases from the New England Journal of Medicine, a modern DDX generator suggested the correct diagnosis 96% of the time[9]. Advances in computer software and hardware have made the new DDX generators far more powerful than earlier programs. Likewise, the ability to integrate more factors in patient presentation, such as geography, demographics, and past diagnoses, makes their suggested diagnosis list more accurate and useful. The most recent developments allow for at least partial integration into the electronic health record (EHR) so that the DDX generator is drawing upon real-time information about the patient and hence requires less manual data entry[10].

Because of these recent advances, the authors felt that a review of the current state of the technology was in order. The findings of this review may help drive research and/or product development agendas. We also seek to highlight the most helpful features along with those barriers and challenges that remain from the perspective of practicing clinicians. The review uses consensus criteria to compare and contrast the DDX generators most relevant to the generalist facing an undiagnosed patient.

## METHODS

Our author group consisted of a medical librarian with expertise in search strategies in evidence-based medicine, and physicians with expertise in computerized decision support, cognitive error, patient safety and education in the diagnostic process. The specialty areas of pediatrics, emergency medicine, and internal medicine were represented in the authorship group. Thus, the perspective was that of generalists faced with undiagnosed patients in the emergency department, inpatient, and office-based settings.

Consensus was achieved on the inclusion and exclusion criteria (listed in Table 1) before the search for DDX generators. The search was conducted in Pubmed and Google® (see online Appendix 1 for details). Clinical decision support systems were defined as "any computer program(s) designed to help healthcare professionals to make clinical decisions."[11] Wyatt & Spiegelhalter included a criterion that such programs "use two or more items of patient data."[12] The authorship team built upon this starting definition through a series of consensus-building meetings via web teleconferencing. We defined DDX generators as programs which assist healthcare professionals in clinical decision making by generating a DDX based on a minimum of two items of patient data. These included signs, symptoms, disease characteristics and/or other patient data.

**Table 1. Inclusion and Exclusion Criteria for DDX Program Review**

| Inclusion Criteria | Exclusion Criteria |
|---|---|
| Diagnosis list | No diagnosis list or static list |
| Ranking of diagnoses or indication of critical diagnoses | No ranking of diagnoses or indication of critical diagnoses |
| Ability to enter/choose at least two symptoms/ characteristics | No ability to enter/choose at least two symptoms/ characteristics |
| Ability to compare diagnoses | No ability to compare diagnoses |
| Intended for heath care providers | Intended for patients |
| Focus on general medicine | Focus on one disease/medical specialty |
| Access via hospital library or physician subscription | Not accessible |

After preliminary review of identified programs, consensus ensued on factors (Table 2) to include in the evaluation round. These factors were then assessed for each DDX generator by two independent evaluators. In cases of disagreement on review criteria, a consensus discussion ensued. When information was not available to the reviewer, the company producing the software was queried for clarification. In cases where we had no response from a vendor on a particular question, and the answer was not clear from publicly available reference materials, we listed the item as unknown.

The evaluation criteria were built upon work by Musen, Shahar and Shortliffe who characterize clinical decision support systems based on five dimensions: the system's intended function, the advice mode, the communication style, the underlying decision-making process and the factors related to human-computer interaction[11]. These criteria were considered and refined through consensus discussion into the evaluation criteria listed in Table 2. The method of inputting data into the system was considered one of the most important criteria, as was the ability to refine the criteria after the initial input. The underlying technique for generating the differential diagnosis by the program was recorded to the degree that the program creator reveals how the program works. Additional features incorporated into the study for descriptive and comparison purposes included: the pricing model, frequency of updating, usage tracking, ability to access further information via references, and other features deemed by the reviewer to be subjectively important. The ability to integrate with the EHR was incorporated as an evaluation criteria, but actual EHR integration was not tested due to resource limitations. With increasing federal emphasis on interoperability of EHRs, adherence to Health Level 7 (HL7) interoperability standards was also considered.

We conducted basic performance testing by entering 20 cases into the four DDX generators. Ten consecutive diagnosis-focused cases chosen from an arbitrary start date were selected from 2010 editions of the Case Records of the New England Journal of Medicine (NEJM)© and from the Medical Knowledge Self Assessment Program (MKSAP)©, version 14, of the American College of Physicians (see online Appendix 3 for case list and scores). Without knowledge of the diagnosis, up to 10 key findings for each case were selected by one of the authors (MLG). These key findings were then entered into the DDX generators by research assistants who were trained to

**Table 2. Evaluation Criteria**

| Criterion | Definition |
|---|---|
| Subscription/ Licensing Model | • Does the vendor offer individual licensing?<br>• Does the vendor offer institutional licensing? |
| EHR Integration | • Can the program be integrated with an EMR? |
| Version | • List the version of the program |
| Health Level 7 (HL7) Interoperability Standards | • Does the program incorporate Health Level 7 (HL7) interoperability standards? |
| Input Elements and Methods | • Does the program pull any data from the EMR to pre-populate fields in the DDX program?<br>• Are sign/symptoms manually entered or selected from list?<br>• Can numeric lab values be entered/pre-populated?<br>• Can negative lab values be considered?<br>• Can patient's current drug list be entered/pre-populated?<br>• Can the program take into account the geographic location/altitude of the patient (e.g. for Rocky Mountain spotted fever, altitude sickness)? |
| Mobile Access | • Does the program provide a mobile device interface? |
| Filtering/Refinement | • Is further filtering/refinement of diagnosis list possible (e.g., by age group, gender, etc.)?<br>• Can patient demographics be pre-populated from EMR? |
| Mechanism of Generating Potential Diagnoses | • What is the ordering of diagnoses based on?<br>• Does the program use natural language processing?<br>• Detail any type of weighting that figures into generating the diagnosis<br>• Does program list drugs that can cause the collection of signs/symptoms? |
| Evidence-Based Content | • Is the content provided evidence-based or incorporate evidence-based guidelines?<br>Detail sources |
| References | • Does the program provide bibliographic/textbook references for the diagnoses presented?<br>• Can it provide links to full text articles? |
| PubMed or other Search on Diagnosis | • Does the program allow for PubMed Linkout to provide access to full text of library/institutionally subscribed resources? |
| Content Source | • What is the source of any drug information provided? |
| Updating | • How often is the content updated? |
| Usage Tracking | • Is it possible to obtain reports on the level of usage of the program? |
| Other features | • Detail any additional features |

enter as many of the findings as the program would allow, and who were also unaware of the final diagnosis until after the searches were conducted. One research assistant entered the case across all four ddx generators to reduce variability in the method of input of the findings. The results generated were then reviewed to see if the correct diagnosis was listed in the first 20 suggestions or the first screen of DiagnosisPro® suggestions (not strictly one page due to formatting) using a 0–5 scoring system:

5 The actual diagnosis was suggested on the first screen or in the first 20 suggestions.
4 The suggestions included something very close, but not exact.
3 The suggestions included something closely related that might have been helpful.
2 The suggestions included something related, but unlikely to be helpful.
0 No suggestions close to the target diagnosis.

In cases where a research assistant was uncertain as to the grading level, the case was discussed with one of the authors (MLG). The use of assigned scores allowed comparison of the results using parametric statistics by analysis of variance with Dunnett T3 correction for multiple comparisons (SPSS© 15.0, Chicago, IL). We also totaled the number of exact matches from each program as an additional marker of performance.

## RESULTS

A total of 23 programs were identified during our initial search. After the application of the exclusion criteria, 11 programs were excluded because of specialty-specific focus (see online Appendix 3 for all excluded programs). Another eight programs were excluded after an initial review for reasons that included: inability to compare diagnoses, inability to enter two symptoms or characteristics, a static tree structure with cross linking of internal reference points, and no ranking of the diagnoses. Four programs were reviewed fully with the evaluation criteria listed in Table 2. The general information for each of the programs is listed in Table 3. Information regarding data elements available for input and input methods are listed in Table 4, and information regarding DDX content sources are listed in Table 5.

Knowledge regarding the mechanism of generating the DDX results is limited to the information shared by the vendors. For DiagnosisPro® the underlying logic was not specified. The diagnoses are presented in disease categories. The results are not rank ordered in terms of disease prevalence or other criteria and the program offers no advice on how to further refine the suggestions. These factors limited the program's usefulness. One differentiating feature is that DiagnosisPro® progressively truncates the list of suggestions as additional findings are entered. Conversely, with the other generators, the lists are re-prioritized, but remain large.

DXPlain® rank ordered results from most to least likely within two categories: common vs. rare diseases, based on disease prevalence. The mechanism is presumed to be a propriety algorithm from the description that follows. An importance rank is given based on criticality of potential diagnosis. Findings are assigned two attributes: one relating to the frequency of the finding in the disorder, and one expressing how strongly it suggests that disease. Ranking is related to findings that are both important and suggestive of a disorder. Rank of a given disease will be lowered if findings commonly seen in the disease are stated to be absent. The attributes are used to generate an ordered list of diagnoses associated with some or all of a given set of findings. Of note,

**Table 3. General Information**

| | Diagnosis Pro® | DXPlain® | Isabel© | PEPID™ |
|---|---|---|---|---|
| Producer | MedTech USA, Inc 6310 San Vincente Blvd. Suite 404, Los Angeles, CA 90048 | Laboratory of Computer Science of the Department of Medicine Massachusetts General Hospital Boston, MA 02114 | Isabel Healthcare Inc. P.O. Box 8393, Reston, VA 20195 | Pepid Medical Information Services LLC, 1840 Oak Ave., Suite 100, Evanston, IL 60201 |
| Subscription/ Licensing Model | Institutional and Individual (free online with advertising) | Institutional only | Institutional and individual | Institutional and Individual. Available as an add-on to PEPID |
| Version | 6.0 | December 18, 2010 | Version 3 | Version 11.1 |
| Health Level 7 (HL7) Interoperability Standards | Unknown | Under development | Yes | Yes |
| Mobile Access | Yes | No | Yes | Yes |
| Updating | Continuously | Monthly except for immediate critical updates | Major updates are performed every 3–6 months | Every 8–10 weeks except for immediate critical updates |
| Usage Tracking (Institutional Subscription/ Licensing) | None mentioned | Yes | Yes | Yes |

DXPlain® allows occupation as a finding, the input of negative findings such as "no fever," and has a side-by-side disease comparison feature. The program displays supportive findings and guides the user to other findings which, if present, support or refute the disease.

Isabel© was the only program to accept natural language queries and the only product allowing the user to input all of the key findings at once. The program uses a "natural language processing" search engine to match entered clinical features with similar terms in the diagnostic data set. Each diagnosis has a complete description of the clinical features with the differential ranked by the strength of the match to the entered clinical features. With each clinical feature addition, the differential diagnostic output reconfigures the list, taking into account all the clinical features entered. Isabel has links to databases, knowledge sources and validation studies.

PEPID™ lists diagnoses based on a proprietary scoring system related to the number of selected signs/symptoms consistent with each potential diagnosis. Additionally, each sign/symptom is assigned a unique score/weight relative to its importance in differentiating among specific diagnoses. Classic or cardinal disorders in which selections strongly suggest a disease or are pathognomonic are indicated. Critical diagnoses with immediate life or limb threat are indicated. Worthy of note is that the overall PEPID™ product, of which the DDX generator is only one piece, incorporates a laboratory testing manual, a drug interactions generator, a drug database covering 7,500 drugs, approximately 400 interactive clinical calculators, an IV compatibility tool, an acute care / life support reference section, and 700 evidence based topics (primary care module).

None of the vendors allowed for unfettered access to institutional library resources or PubMed Linkout for full text from subscribed content, although both Isabel and DxPlain® do provide for Pubmed searching. DiagnosisPro® and Isabel report that they integrate with major EHR vendor products to some degree, but we did not test the ability to integrate any of the products into an EHR. It is noteworthy that DiagnosisPro® has English, French, Spanish, and Chinese interfaces.

Aggregated results and mean scores (with 95% confidence intervals) from entering published cases into each of the differential diagnosis generators are shown in Table 6. ISABEL© and DxPlain® performed well with means of 3.45 for both. Post-hoc analysis with correction for multiple comparisons revealed that only the difference between DxPlain® and PEPID™ reached statistical significance (P=0.04, mean score

**Table 4. Input Elements and Methods**

| | DiagnosisPro® | DXPlain® | Isabel© | PEPID™ |
|---|---|---|---|---|
| Imaging/ Diagnostic Test Results | Yes | Yes | Yes | Yes, for chest x-rays, however, other imaging findings are not supported |
| Patient Demographics | No | Yes | Yes | Yes |
| Can Input Lab Values (e.g. High Potassium) | Yes | Yes | Yes | Yes |
| Can Input Medications | No | Yes, for specific overdoses or class of medication only | Yes | No |
| Can Input Geography | Yes | Yes | Yes | No |
| Negative Findings Considered | No | Yes, a negative finding like absence of fever can be entered by checking "Finding absent" and then searching for that finding | Limited to natural language structure (e.g. absence of fever) | No |
| Information Can Be Populated from EHR | Yes | Currently limited to EHR at Massachusetts General Hospital | Yes | No |

Table 5. Content Sources and Linking

| | DiagnosisPro® | DXPlain® | Isabel© | PEPID™ |
|---|---|---|---|---|
| Content Source | Textbooks, journal articles and websites | Proprietary knowledge base | Federated search of leading texts and journals | Proprietary knowledge base |
| Evidence Based | No | Partial. Specific evidence-based recommendations from specialty societies and CDC considered in content development | Partial. Specific evidence-based recommendations are considered in content development | Partial. Specific evidence-based recommendations and analyses which are incorporated contain graded recommendations from FPIN and BEEM |
| References | 104 references are listed as their sources of content, but specific disease information does not display a specific reference | References to Medline abstracts and open access guidelines. Number of references dependent upon topic | The "knowledge" choice on the tool bar allows a search of approximately 90 journals and 7 online texts | References to evidence-based information from FPIN integrated in primary care module. Other sources are cited throughout. Number of references dependent upon topic |
| PubMed or other Search on Diagnosis | Can run preformatted PubMed search from disease description screen | Can run preformatted PubMed search and/or structured Google® search of pre-selected medical websites | Can run preformatted PubMed search, and search texts, journal abstracts, images, and web resources | No link to PubMed |
| Drug Content Source | Uncertain. Reference list includes many possible sources for drug information | No specific drug information provided | Martindale and other sources | American Society of Hospital Pharmacists |

*Abbreviations BEEM = Best Evidence in Emergency Medicine*
*FPIN = Family Practice Inquiries Network*

difference 1.75, 95% C.I. 0.05 to 3.45) None of the generators included the correct diagnosis for two of the MKSAP cases (acquired von Willebrand's disease related to aortic stenosis, and metformin-induced peripheral neuropathy). Certain scores for returned suggestions such as "pancreatitis" for autoimmune pancreatitis and "cardiomyopathy" for methamphetamine-induced cardiomyopathy were scored only "3" (or "might have been helpful") because the broad category of diagnosis was clear from the presentation and the DDX generator did not help elucidate the root cause. Compared to the three other generators which appeared to have large vocabularies, PEPID™ was unable to recognize many of the key findings. The number of exact matches was DiagnosisPro®=5, DxPlain®=10, Isabel©=9, and PEPID™=4.

## DISCUSSION

This evaluation is intended to raise awareness of the existence of the DDX generators for clinical use and teaching. It also serves as a framework for institutions to use in considering purchase or subscription. Differential diagnostic generators have matured significantly and have begun to leverage access to the EHR, the internet and, to the degree allowed by vendors, subscription-based resources. Potential barriers to the use of DDX generators include access due to subscription models for the generators themselves, lack of the EHR with which to integrate, limitations of the user interface and lack of access to linked content (both EBM and non-EBM). In regard to adoption, we should note that two of the four programs tested by Berner et al.[7]. are no longer sold, and that DXPlain® is not available to the individual physician. Overall, all of the programs put forth for the final review and testing were deemed subjectively assistive and functional for clinical diagnosis and education.

Table 6. Scores for Case Testing

| | Diagnosis Pro® | DXPlain® | Isabel© | PEPID™ |
|---|---|---|---|---|
| MKSAP Case Score (maximum 50) | 22 | 31 | 34 | 22 |
| NEJM Case Score (maximum 50) | 31 | 38 | 35 | 12 |
| Aggregate Score (maximum 100) | 53 | 69 | 69 | 34 |
| Mean Score (0–5) (95% Confidence Interval) | 2.2.65 (1.75, 3.55) | 3.45 (2.63, 4.27) | 3.45 (2.53, 4.37) | 1.70 (0.71, 2.69) |

While DDX generators have been shown to solve very complex cases[9], the question of helpfulness among experts in real time clinical diagnosis remains. The expert goes through a series of hypothesis refinement in complex cases[13], much the way the diagnosis becomes more refined as more items are added to the DDX generator input. Studies are needed to test these systems' ability to render final diagnosis more quickly and to support safety in the diagnostic process without overburdening alarms. Berner et al.[7]. discussed the issue of diagnostic relevancy and the fact that long lists of diagnoses may be unusable by the practicing clinician and challenging for students[18]. This paper did not specifically address the relevancy or length of the diagnostic lists; in addition, the signal-to-noise problem is difficult to avoid in this setting. For example, the progressively truncated lists generated by DiagnosisPro® improve the diagnostic specificity, but at the expense of sensitivity. We share the concern that, especially in novice clinicians, the lists could lead to increased diagnostic testing with concomitant risk for increased costs and/or iatrogenic injury. Such a factor would be very

difficult to quantify in practice. Still, the more urgent consideration is that human memory dictates that the list of diagnoses considered at any one time will be limited, and that the risk of not considering the diagnosis (sensitivity) is the greater concern.

While all of the programs provide a means for manual entry of findings, only two have reached the level of populating this information from various EHRs (Isabel© and DiagnosisPro®). We did not engage in EHR integration testing, which would require fees and an integration process. Also, we did not test the programs in clinical practice with the incumbent workflow and time pressures, something highly recommended prior to purchase or integration decisions. Healthcare systems with significant EHR usage and with single vendor EHRs across multiple settings may find integration more cost effective. We would caution that consideration of whether or not these programs add to an institution's ability to meet "meaningful use" criteria set by the Health Information Technology for Economic and Clinical Health (HITECH) Act was beyond the scope of this evaluation.

Those DDX generators that can be integrated with the EHR are currently limited in their connectivity to the assigned fields shared with the generator. A different strategy would take all potentially relevant data and share it with the DDX generator in real time; new products that take this integrative approach are currently in development and testing (Lifecom©)[14]. In this manner, the DDX generator hypothesis is evolving in real time by updating the known problem set. This may help overcome one of the classic problems of cognitive error—the challenge of knowing when to use decision support. Because errors are made in simple as well as complex cases, if DDX generators are accessed only by active choice in cases known to be diagnostic challenges, then many cognitive diagnostic errors will proceed unmitigated in the current paradigm.

DDX generators can serve as helpful adjuncts in education. Bowen et al. recently described how a detailed review of the learner's DDX using a compare and contrast strategy leads to the development of illness scripts which serve as anchor points in the learner's memory[15]. Students and preceptors alike believe the ability to reflect upon the reasoning process is one of the most valuable pieces of the educational encounter[16,17]. One approach is to have students generate an independent DDX and compare it to the list from a DDX generator. Thus, the preceptor gains insight into the learner's reasoning process and can identify and correct cognitive errors.

None of the programs allow institutions to leverage their current journal subscriptions for full text versions of articles provided in references, although many provide access to PubMed. Vendors should provide a means of allowing institutions to use their library's customized PubMed URL to provide the full text of articles referenced. This linking to EBM resources can seamlessly move the clinician from considering a diagnosis to considering the test, and test properties, for investigating the diagnosis. None of the programs support standard solutions such as the digital object identifier and an open URL link resolver would be another welcome feature.

Limitations of our evaluation include the use of an iterative process rather than a formal Delphi method for achieving consensus regarding inclusion and evaluation criteria. In addition, performance testing was not directed at specificity and comparison of performance between programs was limited in statistical power; however, the results of comparisons using our graded scoring system was very similar to the "exact match" comparison, adding some reassurance as to the validity of our findings that DxPlain® and Isabel© performed the best in identifying the correct diagnosis.

***Contributors:*** *The authors would like to thank medical students Genine Siciliano, Agnes Nambiro, Grace Garey, and Mary Lou Glazer for entering the findings into the DDX generators for testing.*

***Sponsors:*** *This study was not funded by an external sponsor.*

***Presentations:*** *This information was presented in poster format at the Diagnostic Error in Medicine Conference 2010, Toronto, Canada.*

***Conflict of Interest:*** *None disclosed.*

***Corresponding Author:*** *William F. Bond, MD, MS; Department of Emergency Medicine, Lehigh Valley Health Network, 1247 S Cedar Crest Blvd. Suite 202, Allentown, PA 18103, USA(e-mail: william.bond@lvhn.org).*

## REFERENCES

1. CRICO Harvard Risk Management Foundation. High Risk Areas: 26% of claims are in the category of diagnosis. Accessed May 30th, 2011, at http://www.rmf.harvard.edu/high-risk-areas.
2. **Brown TW, McCarthy ML, Kelen GD, Levy F.** An epidemiologic study of closed emergency department malpractice claims in a national database of physician malpractice insurers. Acad Emerg Med. 2010;17(5):553–560.
3. **Croskerry P.** Clinical cognition and diagnostic error: applications of a dual process model of reasoning Advances In Health Sciences Education. Theory And Practice. 2009;14(Suppl 1):27–35.
4. **Schiff, G. D., Kim, S., Abrams, R., Cosby, K., Lambert, B. L., Elstein, A. S., Hasler, S., et al.,** Diagnosing Diagnosis Errors: Lessons from a Multi-institutional Collaborative Project. Advances in Patient Safety: From Research to Implementation. Volumes 2, AHRQ Publication Nos. 050021 (Vols 1–4). February 2005. Agency for Healthcare Research and Quality, Rockville, MD. Accessed May 30, 2011, at http://www.ahrq.gov/qual/advances/.
5. **Schiff GD, Bates DW.** Can Electronic Clinical Documentation Help Prevent Diagnostic Errors? N Engl J Med. 2010;362(12):1066–1069.
6. **Barnett GO, Cimino JJ, Hupp JA, Hoffer EP.** DXplain An evolving diagnostic decision-support system. JAMA. 1987;258(1):67–74.
7. **Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, Ball EV, et al.** Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330(25):1792–1796.
8. **Kassirer JP.** A report card on computer-assisted diagnosis–the grade: C. N Engl J Med. 1994;330(25):1824–1825.
9. **Graber ML, Mathew A.** Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med. 2008;23(Suppl 1):37–40.
10. **Ramnarayan P, Cronje N, Brown R, Negus R, Coode B, Moss P, Hassan T, et al.** Validation of a diagnostic reminder system in emergency medicine: a multi-centre study. Emerg Med J. 2007;24(9):619–624.
11. **Musen, M. A., Shahar, Y. and Shortliffe, E. H.,** Clinical Decision Support Systems, In: Shortliffe, E. H. and Cimino, J. J. (eds), Biomedical

Informatics: Computer Applications in Health Care and New York: Springer, 2006, pp. 698–736.
12. **Wyatt, J. and Spiegelhalter, D.,** Field trials of medical decision-aids: potential problems and solutions. , Proceedings of the Annual Symposium on Computer Application in Medical Care, 1991, pp. 3–7.
13. **Kassirer J, Wong J, Kopelman R.** Learning Clinical Reasoning. Philadelphia: Lippincott Williams and Wilkins; 2009.
14. **Datena, S.,** Lifecom DARES Approach to Problem Based Learning and Improvement (Personal Communication of Unpublished White Paper describing the Lifecom DARES System), 2010.
15. **Bowen JL.** Educational strategies to promote clinical diagnostic reasoning. N Engl J Med. 2006;355(21):2217–2225.
16. **Wolpaw T, Papp KK, Bordage G.** Using SNAPPS to facilitate the expression of clinical reasoning and uncertainties: a randomized comparison group trial. Acad Med. 2009;84(4):517–524.
17. **O'Malley PG, Kroenke K, Ritter J, Dy N, Pangaro L.** What learners and teachers value most in ambulatory educational encounters: a prospective, qualitative study. Acad Med. 1999;74(2):186–191.
18. **Graber ML, Tompkins D, Holland JJ.** Resources medical students use to derive a differential diagnosis. Med Teach. 2009;31(6):522–527.