OPEN ACCESS

Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel

Matt Sibbald ,[1] Sandra Monteiro ,[2] Jonathan Sherbino,[1] Andrew LoGiudice,[3] Charles Friedman,[4] Geoffrey Norman[2]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2021-013493).

[1]Department of Medicine, McMaster University, Hamilton, Ontario, Canada
[2]Department of Health Evidence and Impact, McMaster University, Hamilton, Ontario, Canada
[3]McMaster University, Hamilton, Ontario, Canada
[4]University of Michigan, Ann Arbor, Michigan, USA

Correspondence to
Dr Matt Sibbald, Department of Medicine, McMaster University, Hamilton, ON L8N 3Z5, Canada; sibbald@mcmaster.ca

Received 9 April 2021
Accepted 9 September 2021

To cite: Sibbald M, Monteiro S, Sherbino J, *et al*. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2021-013493

ABSTRACT

Background Diagnostic errors unfortunately remain common. Electronic differential diagnostic support (EDS) systems may help, but it is unclear when and how they ought to be integrated into the diagnostic process.

Objective To explore how much EDS improves diagnostic accuracy, and whether EDS should be used early or late in the diagnostic process.

Setting 6 Canadian medical schools. A volunteer sample of 67 medical students, 62 residents in internal medicine or emergency medicine, and 61 practising internists or emergency medicine physicians were recruited in May through June 2020.

Intervention Participants were randomised to make use of EDS either early (after the chief complaint) or late (after the complete history and physical is available) in the diagnostic process while solving each of 16 written cases. For each case, we measured the number of diagnoses proposed in the differential diagnosis and how often the correct diagnosis was present within the differential.

Results EDS increased the number of diagnostic hypotheses by 2.32 (95% CI 2.10 to 2.49) when used early in the process and 0.89 (95% CI 0.69 to 1.10) when used late in the process (both p<0.001). Both early and late use of EDS increased the likelihood of the correct diagnosis being present in the differential (7% and 8%, respectively, both p<0.001). Whereas early use increased the number of diagnostic hypotheses (most notably for students and residents), late use increased the likelihood of the correct diagnosis being present in the differential regardless of one's experience level.

Conclusions and relevance EDS increased the number of diagnostic hypotheses and the likelihood of the correct diagnosis appearing in the differential, and these effects persisted irrespective of whether EDS was used early or late in the diagnostic process.

BACKGROUND

The need for improved accuracy and rapidity of medical diagnoses—extensively documented in the literature[1–5]—became an early focus of the field of medical informatics.

Early efforts to augment the diagnostic process with information technology in the 1970s, using the methods of artificial intelligence and so-called 'expert' systems that were available at the time, addressed only specific domains such as infectious[6] and gastrointestinal[7] diseases. These efforts were followed in the 1980s and 90s by more sophisticated and broadly applicable electronic diagnostic support (EDS) systems that spanned most of internal medicine—QMR, ILIAD, DxPLAIN—among others.[8–11]

These EDS systems were studied from multiple perspectives. Some studies addressed the accuracy of the systems themselves. That is, when data were entered from a case with a known diagnosis, how often did the system yield the correct diagnosis?[12] Other studies examined how well the systems augmented the diagnostic reasoning of clinicians across the spectrum of training and experience when confronted with diagnostically challenging cases.[13] That is, do clinicians of varying experience levels make more accurate diagnoses with the aid of EDS than without? Individual studies and meta-analyses revealed small but statistically significant improvements in diagnostic accuracy.[8–11 14 15] However, these systems required user training and time-intensive manual entry of data, limiting EDS system use to selected cases rather than routine, regular use. Possibly because of their only modest improvements to diagnostic accuracy,[8–11 14 15] time-consuming manual data entry[13] or

clinician perceptions that they are unneeded,[1 16 17] many of these systems from the late 20th century have largely passed into obscurity.

A newer system, Isabel, provides a list of relevant diagnoses via a user-friendly, time-efficient platform. The system is designed around entering a small number of symptoms in free-text without the need for symptom qualifiers, pertinent negatives, medical, social or family history background, physical signs, lab values or investigations. Generating differential diagnoses in this way dramatically reduces the time required to several minutes, making it feasible at point-of-care for both physicians and patients.[18–23] Though there is no publicly available technical information on the computational methods by which Isabel generates diagnostic hypotheses, the tool is advertised as using artificial intelligence and natural language-processing techniques, and has been adopted in both experimental[18–21] and practical[22 23] settings. When Isabel was used by clinicians admitting paediatric patients to the intensive care unit, a 4% improvement in diagnostic accuracy was documented—representing a 30% reduction in harm.[17] And so while the improvement in diagnostic accuracy may seem small, it is nonetheless important.

Prior studies of Isabel speak to the accuracy of diagnostic hypotheses it generates, but its potential to augment a clinician's own diagnostic reasoning remains unclear.[24 25] In particular, it is unknown whether Isabel should be integrated early in the diagnostic process during the hypothesis generation stage, or later during the deductive analytical stage.[25] The present study addresses this gap by replicating and building on a comprehensive study of ILIAD and QMR published more than 20 years ago.[13] Using a similar method and identical clinical test cases, here we examined whether Isabel brings clinical medicine closer to the full potential of diagnostic decision support.

To elucidate how Isabel can be best integrated, we explored whether its benefits are moderated by early or late implementation in the diagnostic process. Early use may increase its effectiveness because all data have not yet been gathered and hypotheses are still formative, so it has the potential to shape the collection and interpretation of data during a history or physical. Conversely, its use later in the encounter when all data are available may provide a final check for unconsidered diagnoses. We therefore designed a study to test these two interventions: early EDS use with only the patient demographics and chief complaint, and late EDS use with all available clinical details. We hypothesised that early EDS use would increase the number of differential diagnostic hypotheses generated, but, because available data are limited, will have minimal impact on accuracy of diagnosis. Conversely, we hypothesised late EDS use would improve diagnostic accuracy, but minimally impact hypothesis generation.

METHODS

We explored the impact of Isabel by asking clinicians of varying levels of expertise to work through a series of cases on an online web-based platform, providing a differential diagnosis before and after the use of the Isabel EDS and randomising them to receive EDS either early (when limited information was available) or late (when all information was available) in the diagnostic process. We measured the number of diagnostic hypotheses recorded, presence of the correct diagnosis within the differential and time spent working through the cases.

Design and procedures

Participants were randomised into one of two groups: early or late use of EDS using a standard random number generator in Excel (Microsoft, Redmond). Clinicians in the early group were randomised to use the EDS after only the patient's demographics and chief complaint were presented, whereas those in the late group used the EDS system after all case details were presented. The entire study was administered via an online platform.

A schematic of the design is shown in figure 1. Participants were told they would be diagnosing written clinical vignettes using an EDS system at some predetermined point in the process. For each case they first received the patient's demographics and chief complaint (eg, a 32-year-old man presenting with diplopia and difficulty swallowing) and provided an initial differential diagnosis, including as many diagnoses as they felt relevant in decreasing order of likelihood.

The two groups differed only after this provision of the initial differential. Participants in the early group were then asked to (1) use the EDS, then provide a revised differential diagnosis, then (2) read the rest of the case material and provide a second revised differential. This format assessed the implementation of EDS during the preliminary hypothesis generation stage before the full case information was available. In contrast, participants in the late group (1) read the rest of the case and revised their final differential without EDS, then (2) were given access to EDS and asked to further revise their differential. This late implementation occurred during the analytical ratification or deductive stage that occurs once all the case information is available. It should be noted that participants could not go back or use the software multiple times; they could only access it at the specified time. Thus, both groups generated a differential both before and after use of EDS, but the timing of EDS served to enhance either early hypothesis generation or the deductive process that follows.

Participants

We recruited participants from six Canadian medical schools between 1 May and 1 June 2020. Medical

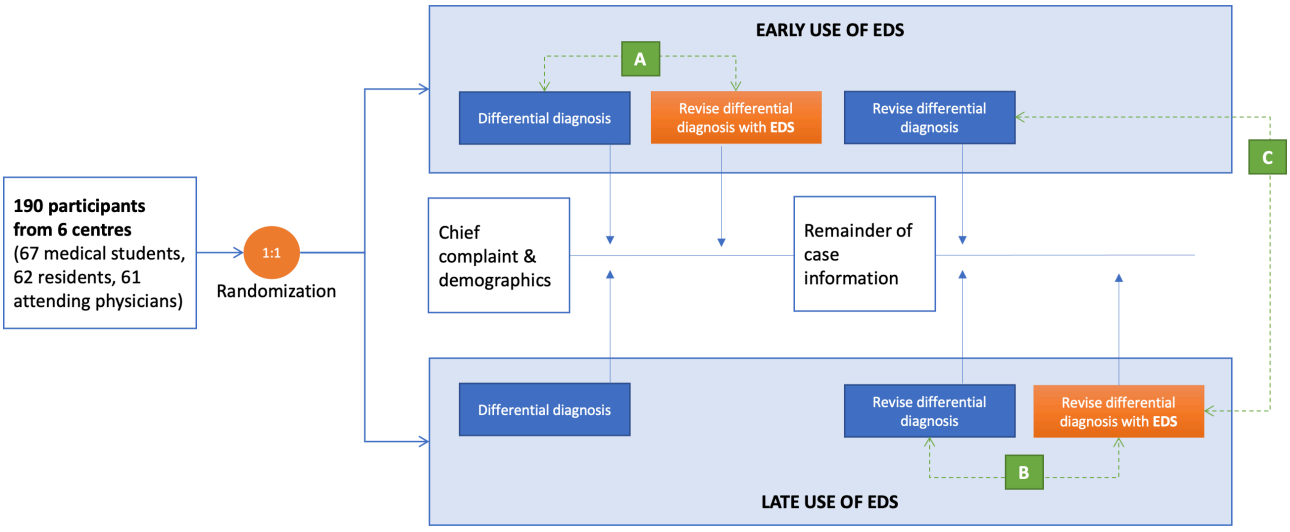

Figure 1 Study design and randomisation. Participants from three centres were randomised to either early or late use of EDS to work through 16 written cases. Both groups provided a differential diagnosis three times. The early group was asked to provide a differential diagnosis after the chief complaint and demographics, revise it with use of EDS and conduct a final revision with the remainder of the case material available. The late group was asked to provide a differential diagnosis after the chief complaint and demographics, revise it with the remainder of the case material and then use EDS to revise a second time. EDS, electronic diagnostic support.

students, internal medicine residents, emergency medicine residents, internists and emergency medicine physicians were invited to participate via email. Participation was voluntary. A modest stipend was provided after participants completed the study (students $125, residents $150, practising physicians $250).

Case materials

Case materials were derived from a previous study involving EDS.[13] The author of the previous study (CF) has maintained control of distribution, so cases cannot have been seen by participants. All cases were based on real patients. Cases were reviewed by two investigators to ensure they were current. Sixteen of the 36 cases available were chosen. We selected the eight easiest and eight hardest cases based on actual performance data from a previous study[13] (see online supplemental appendix 1 for a sample easy and hard case). The EDS was not previously trained on the case material (personal communication, IsabelHealthCare.com).

Electronic differential diagnosis support

In order to use the EDS, clinicians enter patient demographics and a list of symptoms into an online platform (IsabelHealthCare.com)[24] which generates a list of diagnoses to consider, annotating diagnoses that ought not be missed (figure 2). We verified that Isabel contained the correct diagnosis of all the test cases in its database.

Participants completed a sample case to familiarise themselves with the EDS prior to the study. Pilot work indicated that no formal training was required, as even non-medical volunteers were able to use the interface without receiving any instructions.

Outcomes

Our two principal outcomes of interest were: (1) the number of diagnoses proposed by the clinician per case and (2) presence of the correct diagnosis within the differential. The latter measure involved a binary approach, whereby the correct diagnosis (or its synonyms) being present anywhere on the list was scored as a 1 or its absence as a 0. Lists of synonyms were available from a previous study.[13] For each participant, an average was calculated for both outcomes across all 16 cases. Consistent with prior work, we also scored diagnostic accuracy with two additional algorithms: a score that captures the order of the correct diagnosis in the differential as well as the presence of the correct diagnosis in the first seven items listed.[13] Analysis of these approaches led to the same conclusions as the simple presence or absence score, and so they have been omitted.

We also timed each step during the diagnostic process: (1) time taken to formulate a differential based on the chief complaint, (2) time taken to revise the differential given all the case materials and (3) time taken to use the EDS.

Analysis

The analysis compared differences between participants using the EDS early or late in the diagnostic process, and across different levels of experience. The primary comparisons are shown in figure 1 labelled A, B and C. SPSS version 26 (IBM) was used for all comparisons.

Comparison A focuses on the effectiveness of EDS for the early group, examining differences between the initial differential (based on demographics and chief complaint) and a subsequent revised differential

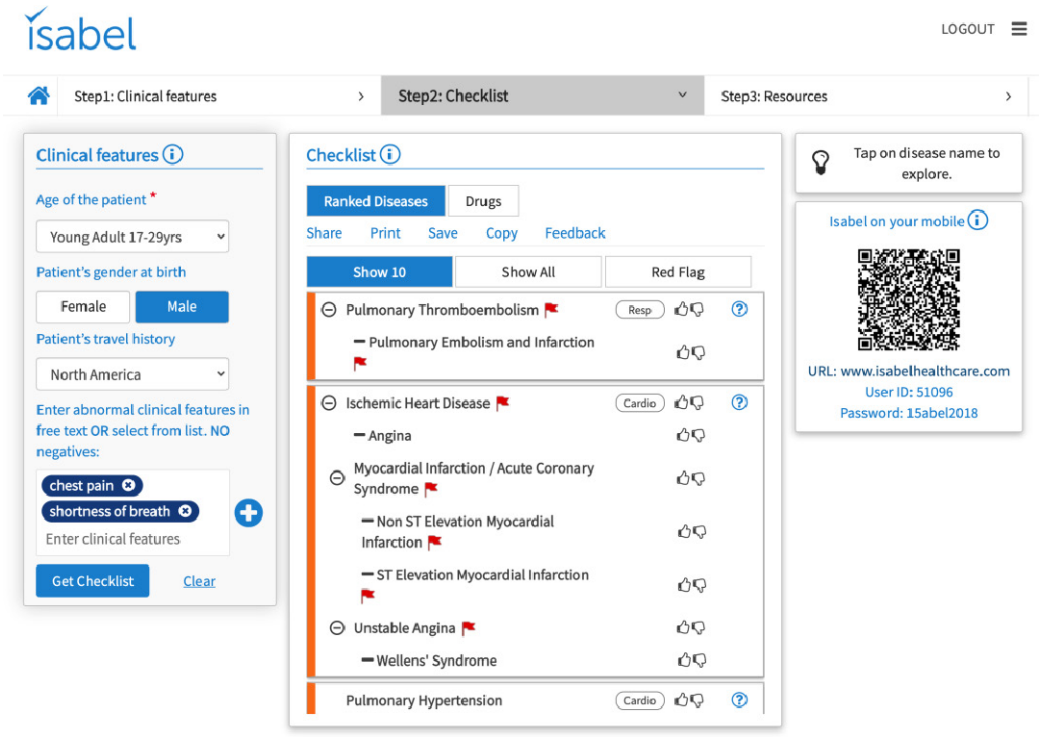

Figure 2 Isabel interface.

following use of EDS. Comparison B focuses on the effectiveness of EDS for the late group, examining differences between a later differential (based on all case information) and a subsequent revised differential following use of EDS. For each comparison, we performed separate repeated-measures analysis of variance (ANOVA) for the two outcomes (average number of hypotheses across all 16 cases, average diagnostic accuracy across all 16 cases) using the within-subject factor of time point (before EDS vs after EDS) and the between-subject factor of experience level (student vs resident vs practising physician).

Comparison C instead focuses on overall differences between the early group and late group by comparing their final differentials. Here we performed separate ANOVAs for the two outcomes (average number of hypotheses, diagnostic accuracy) using group (early vs late) and experience level (student vs resident vs practising physician) as a between-subject factors, and case difficulty (easy vs hard) as a within-subject factor.

Finally, we compared the average time spent using the EDS by participant between the early versus late groups using an unpaired t-test.

Effect sizes were expressed as partial eta squared, η^2, which represents the ratio of the variance attributed to the effect compared with all measured variance. Values of 0.01, 0.06, and 0.13 were interpreted as small, medium, and large effects.[26] Using the Bonferroni correction for multiple statistical comparisons, a p value of <0.008 would result in rejection of the null hypothesis accepting an overall type I error of 0.95. All interactions were considered hypothesis generating rather than confirmatory[27 28] using an alpha criterion of 0.05.

We performed a sample size calculation based on prior published data,[13] anticipating an effect size of 0.35 for the overall benefit of EDS with respect to diagnostic accuracy. Using a conventional alpha of 0.05 and beta of 0.20, this equates to a sample size of 130 to detect a main effect of EDS. However, differences between early and late EDS use have no evidence on which to base a sample size calculation. Presuming that a relevant difference in early or late use of EDS would have to be quite substantial to meaningfully impact practice, we used an effect size of 0.50 to calculate a sample size of 64 per group using the conventional alpha and beta values. SPSS version V.26 (IBM, Redmond) was used.

RESULTS

One hundred and ninety participants across six institutions took part in the study. Participants included 67 medical students, 62 residents in internal medicine or emergency medicine, and 61 practising internists or emergency medicine physicians. For the initial differential (ie, before participants from either group received EDS), there was no difference between the early and late groups in the number of hypotheses proposed, nor the likelihood of the correct diagnosis being present (see table 1). At this point, only 10%–13% of participants identified the correct diagnosis.

Table 1 Baseline characteristics and performance of participants randomised to the early and late use of EDS groups

	Early use of EDS (n=99)	Late use of EDS (n=91)
Female gender	44 (44%)	35 (38%)
Expertise level		
Medical students	37	30
Residents	31	31
Practising physicians	31	30
Centre		
McMaster University	66	56
Northern Ontario School of Medicine	1	0
University of Ottawa	14	16
University of Saskatchewan	7	7
University of Toronto	1	1
Western University	9	8
Not specified	1	3
Baseline performance		
Number of hypotheses generated	5.64±1.79	5.05±1.76
Likelihood of correct diagnosis in differential	0.12±0.06	0.11±0.06

EDS, electronic diagnostic support.

Comparison A: effectiveness of the EDS early in the diagnostic process

Early EDS use increased the average number of hypotheses across all experience levels from 5.64 to 7.96 (mean difference 2.32, 95% CI 2.10 to 2.49). Students showed the greatest gain, from 4.67 to 7.39 (mean difference 2.72, 95% CI 2.40 to 3.03); residents from 6.47 to 8.89 (mean difference 2.42, 95% CI 2.08 to 2.77) and practising physicians from 5.97 to 7.71 (mean difference 1.74, 95% CI 1.40 to 2.09) as shown in table 2. The main effect of time (ie, before vs after EDS) was significant ($F=402$, $\eta^2=0.81$, $p<0.0001$), as was the interaction between time and experience ($F=6.35$, $\eta^2=0.12$, $p=0.003$).

Early EDS use also improved the likelihood of the correct diagnosis being present in the differential from 0.12 to 0.19 across all experience levels (mean difference 0.07, 95% CI 0.06 to 0.09), with improvements from 0.10 to 0.21 for students (mean difference 0.11, 95% CI 0.08 to 0.13), 0.14 to 0.19 for residents (mean difference 0.05, 95% CI 0.03 to 0.08) and 0.12 to 0.17 for practising physicians (mean difference 0.05, 95% CI 0.03 to 0.08). The main effect of time was significant ($F=110.9$, $\eta^2=0.54$, $p<0.0001$), as was its interaction with experience level ($F=7.71$, $\eta^2=0.14$, $p=0.001$).

Comparison B: effectiveness of the EDS late in the diagnostic process

Late EDS use increased the average number of hypotheses from 5.30 to 6.19 (mean difference 0.89, 95% CI 0.69 to 1.10; $F=117.4$, $\eta^2=0.572$, $p<0.0001$), and did not vary by experience level.

Late EDS use also increased the likelihood of the correct diagnosis being present in the differential from 0.21 to 0.30 (mean difference 0.09, 95% CI 0.07 to 0.10), with a significant main effect of time ($F=112.1$, $\eta^2=0.56$, $p<0.0001$), significant main effect of experience level ($F=6.19$, $\eta^2=0.12$, $p=0.003$), and a significant interaction between time and experience level ($F=8.06$, $\eta^2=0.16$, $p=0.001$). Students improved from 0.13 to 0.26 (mean difference 0.13, 95% CI 0.10

Table 2 Outcomes of participants randomised to the early and late use of EDS groups

	Early EDS group			Late EDS group		
	Initial differential	Revision with EDS	Revision with case material	Initial differential	Revision with case material	Revision with EDS
Number of hypotheses generated						
All participants	5.64 (3 to 8.94)	7.96 (4.25 to 11.5)	7.06 (3.44 to 12)	5.05 (2.75 to 8.5)	5.30 (2.25 to 9.94)	6.19 (2.56 to 11.44)
Medical students	4.67 (2.75 to 8.94)	7.39 (4.25 to 11.56)	6.82 (3.25 to 11.75)	4.57 (2.19 to 7.25)	4.8 (1.5 to 8)	5.91 (2.44 to 10.94)
Residents	6.47 (3.88 to 8.75)	8.89 (4.31 to 11.88)	7.92 (3.56 to 12.19)	5.01 (3.19 to 8.31)	5.15 (2.75 to 9.56)	5.86 (3.06 to 11)
Practising physicians	5.97 (4.06 to 9)	7.72 (4.06 to 11.38)	6.48 (3.38 to 11.44)	5.57 (3 to 9.13)	5.95 (2.56 to 11)	6.82 (2.88 to 12.94)
Likelihood of correct diagnosis in the differential						
All participants	0.12 (0 to 0.19)	0.19 (0.13 to 0.25)	0.25 (0.13 to 0.44)	0.11 (0 to 0.19)	0.21 (0 to 0.44)	0.30 (0.06 to 0.50)
Medical students	0.10 (0 to 0.19)	0.21 (0.13 to 0.31)	0.24 (0.13 to 0.38)	0.10 (0 to 0.19)	0.13 (0 to 0.25)	0.26 (0.06 to 0.44)
Residents	0.14 (0.06 to 0.25)	0.19 (0.13 to 0.25)	0.27 (0.13 to 0.50)	0.12 (0.06 to 0.19)	0.25 (0.06 to 0.44)	0.34 (0.19 to 0.56)
Practising physicians	0.12 (0.06 to 0.19)	0.17 (0.06 to 0.25)	0.25 (0.13 to 0.38)	0.12 (0.0 to 0.25)	0.25 (0.06 to 0.44)	0.30 (0.06 to 0.56)

All numbers are means across all 16 cases with 95% CIs shown in brackets.
EDS, electronic diagnostic support.

to 0.15), residents from 0.25 to 0.34 (mean difference 0.08, 95% CI 0.06 to 0.11) and physicians from 0.25 to 0.30 (mean difference 0.05, 95% CI 0.02 to 0.07).

Comparison C: effectiveness of early versus late use of EDS

At the final time point, there was a trend toward more hypotheses being considered by the early group than the late group (7.06 vs 6.19; $F=5.39$, $\eta^2=0.028$, $p=0.02$). This benefit was present for students and residents, but not for practising physicians (interaction with experience level: $F=3.32$, $\eta^2=0.04$, $p=0.04$). However, the correct diagnosis was more likely to be present in the late group than in the early group (0.30 vs 0.25 mean difference 0.05 95% CI 0.01 to 0.08, $F=7.71$, $\eta^2=0.040$, $p=0.006$). The interaction with experience level was not significant.

The main effect of hard versus easy cases on accuracy was significant (hard=0.12, easy=0.44; $F=841$, $\eta^2=0.82$, $p<0.00001$). There was also a small interaction with expertise ($F=10.87$, $\eta^2=0.11$, $p<0.0001$) such that, where students had greater difficulty with hard cases, but performed similarly to more experienced physicians on easy cases (data not shown).

Comparison of early and late groups on time spent using Isabel

Collapsing across both groups, participants took an average of 99 s (95% CI 67 to 130) per case to review the chief complaint and formulate a differential diagnosis, then an additional 188 s (95% CI 173 to 201) to revise this differential upon receiving the rest of the case information. As for time spent using the EDS to further revise their differential, participants took an additional 137 s (95% CI 128 to 148), but these times varied between the early and late groups: participants spent less time with the EDS when using it early in the process compared with late in the process (89 vs 189 s, mean difference 100 s, 95% CI 81 to 124, $p<0.002$).

DISCUSSION

In this multicentre, randomised study, EDS improved the diagnostic process by both increasing the number of diagnostic hypotheses and the likelihood of the diagnosis being present in the differential. The effect size of EDS for both outcomes was large. While clinicians of all experience levels benefited, the impact was most pronounced among novice clinicians.

The effect of EDS on the diagnostic process was present regardless of whether it was used early or late in the diagnostic process. Timing of use did moderate the increases in number of diagnostic hypotheses and likelihood of the correct diagnosis being present. Early use of EDS resulted in an increase of more than twice as many hypotheses as late use (2.32 vs 0.89). Conversely, late use resulted in greater accuracy (ie, inclusion of the correct diagnosis) at the conclusion of the case (0.30 vs 0.25). Such findings suggest that early use of EDS serves to expand the pool of hypotheses under consideration, whereas late use of EDS is useful in helping clinicians carefully decide among alternatives. Ours is not the first study to examine differences in diagnostic support system effectiveness based on when it is used in the diagnostic process. Prior work showed improved diagnostic accuracy with use of a computerised diagnostic support system after a chief complaint and pertinent case details were provided, but not after participants had already proposed a diagnosis.[29] Importantly, our study differed in asking participants to refine a differential rather than establishing and revising a diagnosis, a subtle but important distinction that may account for differences in effectiveness with late use of diagnostic support.

The increased rate of listing the correct diagnosis with EDS in this study is consistent with prior reports using other computer-based diagnostic systems.[13 29] Friedman *et al* observed an improvement of 8% with the QMR EDS and 4% with the ILIAD EDS.[13] Kostopoulou *et al*[29] found a 6% increase with the DxPLAIN EDS when used early in the diagnostic process. These are comparable with the 7% (early) and 8% (late use) improvements in listing the correct diagnosis seen with the Isabel EDS in this study. The larger improvement with novices was also noted in Friedman's work with improvement in diagnostic accuracy for medical students, residents, and practising physicians of 9%, 5%, and 3%, compared with 11%, 5%, and 5% in the present study. In routine clinical practice, diagnostic error rates are estimated to be 1%–15% of all patient encounters.[2 4 5 30] Improvements in diagnostic accuracy as small as 4% resulted in substantial harm reduction,[31] suggesting that if the 5%–11% improvements in the diagnostic process noted in this study were transferable to routine practice, it would be meaningful.

Nevertheless, the critical difference between Isabel and the previous systems is ease of use. In the present study, average time spent with Isabel ranged from 1.5 to 3 min, whereas average times for previous systems were much longer, ranging from 22 to 240 min.[13 32] This is not simply due to a more streamlined interface; unlike prior systems, Isabel uses a minimal amount of (primarily) historical data and yet achieves similar accuracy[18–23] as previous systems that required far more information.[13 29] An important implication of this efficiency is that it becomes feasible to integrate EDS into one's processing of a case in real time.

Interestingly, different advantages with an EDS system at early and late time points are consistent with dual-process models of the diagnostic process from the fields of cognitive psychology[25] and clinical reasoning.[33] Elstein's hypothetico-deductive model proposes that clinicians initially put forth hypotheses based on patient cues, then later ratify them via analytical reasoning.[25] But less experienced clinicians are less likely to generate correct hypotheses[34] and more easily swayed by early diagnostic suggestions[35–37]—a

potential explanation for the disproportionate increase in diagnostic hypotheses adopted by less experienced clinicians with early EDS use. And though more experienced clinicians generate more correct hypotheses, they are still capable of diagnostic error,[33] so late use of EDS may help them reconsider a common diagnosis that was initially overlooked—thereby functioning as a diagnostic checklist[37 38] or a way to re-evaluate their initial impressions.[39] EDS systems can help clinicians guard against search satisficing, where the first diagnosis that comes to mind is too easily accepted[31] and a differential diagnosis is not generated at all.[40] Even experienced clinicians may not recognise a correct diagnosis from the EDS list if they are unfamiliar with it or do not see its relevance to the case. This provides an upper bound on the utility of an EDS in practice, and an important conceptual limit in the use of EDS to improve clinician diagnostic reasoning.

Several limitations of the present study are worth mentioning. First, we used archived cases. While these were derived from actual patient presentations, the codifying of information into written form removes important steps from the authentic clinical process (eg, building a relationship with a patient, using a shared language to elicit symptoms, then interpreting this experience and integrating into a diagnostic process). This complex task competes with the use of EDS in the clinical environment, making it more difficult to use in practice than in studies. Second, whether or not adding a diagnosis to a differential will result in meaningful avoidance of misdiagnosis is still reliant on downstream clinician behaviour to appropriately investigate feasible alternative diagnoses. Clinicians may avoid investigating because of resource constraints, perceived liability or norming pressures within clinical practice groups. These influences may impair some of the benefits of an EDS-assisted approach. Future research should employ EDS in an authentic clinical environment to demonstrate feasibility in the process of care, and also assess whether this benefit with written cases persists in more authentic clinical environments. Third, we might be faulted for our choice of cases, which were designed to be difficult. In prior studies, even academic internists achieved only 40%–50% accuracy.[13] We make no pretence that they are representatives of all problems in practice; quite the opposite. By selecting difficult cases, these are likely more representative of the kind of cases where physicians may seek help from an EDS. Finally, we did not involve a control group tasked with revisiting the differential diagnosis without EDS, leading to an upper limit estimate of EDS effectiveness that includes whatever benefit might be conferred by simply revisiting the differential diagnosis without EDS.

In conclusion, this study demonstrates that the integration of electronic decision support into the clinical diagnostic process across a wide range of experience increases the length of differentials and increases the likelihood of the diagnosis appearing within the differential. This benefit was observed regardless of whether the EDS system was used early or late in the process. In contrast to previous systems, current generation EDS systems are easy and quick to use, permitting easy integration into the clinical reasoning process in real time. Future research should investigate whether these benefits transfer to real-time clinical use.

Twitter Matt Sibbald @sibbaldmatt and Sandra Monteiro @monteiro_meded

Acknowledgements The authors acknowledge Amy Keuhl for her role in project management, Mark Lee for his role in project management and manuscript preparation, Betty Howe for her role in platform design and recruitment.

Contributors All authors contributed to study design, analysis, interpretation and manuscript preparation. MS wrote the first draft which was critically revised by all authors.

Funding This project was funded by a Physicians' Services Incorporated Medical Education Research grant.

Competing interests None declared.

Patient consent for publication Obtained.

Ethics approval Ethical approval was obtained through the Hamilton Integrated Research Ethics Board, protocol #4945. All participants provided written informed consent. Access to Isabel was provided by Isabel HealthCare. The authors have no other relationship and no reporting requirements to the company.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iDs
Matt Sibbald http://orcid.org/0000-0002-0022-2370
Sandra Monteiro http://orcid.org/0000-0001-8723-5942

REFERENCES

1 Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med* 2008;121:S2–23.
2 Graber ML. The incidence of diagnostic error in medicine. *BMJ Qual Saf* 2013;22 Suppl 2:ii21–7.
3 Newman-Toker DE, Makary MA. Measuring diagnostic errors in primary care: the first step on a path forward. Comment on "Types and origins of diagnostic errors in primary care settings". *JAMA Intern Med* 2013;173:425–6.

4 Kwan JL, Lo L, Ferguson J, *et al*. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. *BMJ* 2020;370:m3216.
5 Gunderson CG, Bilan VP, Holleck JL, *et al*. Prevalence of harmful diagnostic errors in hospitalised adults: a systematic review and meta-analysis. *BMJ Qual Saf* 2020;29:1008–18.
6 Shortliffe E. *Computer-based medical consultations: MYCIN*. New York: Elsevier, 2012.
7 De Dombal FT, Leaper DJ, Horrocks JC, *et al*. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians. *Br Med J* 1974;1:376–80.
8 Barnett GO, Cimino JJ, Hupp JA, *et al*. DXplain. An evolving diagnostic decision-support system. *JAMA* 1987;258:67–74.
9 Miller R, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. *MD Comput* 1986;3:34–8.
10 Miller RA, Pople HE, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med* 1982;307:468–76.
11 Warner HR. Iliad: moving medical decision-making into new frontiers. *Methods Inf Med* 1989;28:370–2.
12 Berner ES, Webster GD, Shugerman AA, *et al*. Performance of four computer-based diagnostic systems. *N Engl J Med* 1994;330:1792–6.
13 Friedman CP, Elstein AS, Wolf FM, *et al*. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. *JAMA* 1999;282:1851–6.
14 Johnston ME, Langton KB, Haynes RB, *et al*. A critical appraisal of research on the effects of computer-based decision support systems on clinician performance and patient outcomes. *Ann Intern Med* 1994;120:135–42.
15 Shea S, DuMouchel W, Bahamonde L. A meta-analysis of 16 randomized controlled trials to evaluate computer-based clinical reminder systems for preventive care in the ambulatory setting. *J Am Med Inform Assoc* 1996;3:399–409.
16 Podbregar M, Voga G, Krivec B, *et al*. Should we confirm our clinical diagnostic certainty by autopsies? *Intensive Care Med* 2001;27:1750–5.
17 Meyer AND, Payne VL, Meeks DW, *et al*. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Intern Med* 2013;173:1952–8.
18 Ramnarayan P, Tomlinson A, Rao A, *et al*. ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation. *Arch Dis Child* 2003;88:408–13.
19 Ramnarayan P, Tomlinson A, Kulkarni G, *et al*. A novel diagnostic aid (ISABEL): development and preliminary evaluation of clinical performance. *Stud Health Technol Inform* 2004;107:1091–5.
20 Ramnarayan P, Kulkarni G, Tomlinson A. ISABEL: a novel internet-delivered clinical decision support system. In: *Current perspectives in healthcare computing*, 2004: 245–56.
21 Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. *J Gen Intern Med* 2008;23:37–40.
22 Ramnarayan P, Roberts GC, Coren M, *et al*. Assessment of the potential impact of a reminder system on the reduction of diagnostic errors: a quasi-experimental study. *BMC Med Inform Decis Mak* 2006;6:1–6.
23 Bavdekar SB, Pawar M. Evaluation of an Internet delivered pediatric diagnosis support system (ISABEL) in a tertiary care center in India. *Indian Pediatr* 2005;42:1086.
24 Riches N, Panagioti M, Alam R, *et al*. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. *PLoS One* 2016;11:e0148991.
25 Elstein AS, Shulman LS, Sprafka SA. *Medical problem solving: an analysis of clinical Reasoning*. London: Harvard University Press, 2013.
26 Richardson JTE. Eta squared and partial ETA squared as measures of effect size in educational research. *Educ Res Rev* 2011;6:135–47.
27 Streiner DL, Norman GR. Correction for multiple testing: is there a resolution? *Chest* 2011;140:16–18.
28 Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* 2014;34:502–8.
29 Kostopoulou O, Rosen A, Round T, *et al*. Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. *Br J Gen Pract* 2015;65:e49–54.
30 Singh H, Giardina TD, Meyer AND, *et al*. Types and origins of diagnostic errors in primary care settings. *JAMA Intern Med* 2013;173:418–25.
31 Thomas NJ, Ramnarayan P, Bell MJ, *et al*. An international assessment of a web-based diagnostic tool in critically ill children. *Technol Health Care* 2008;16:103–10.
32 Bankowitz RA, McNeil MA, Challinor SM, *et al*. A computer-assisted medical diagnostic consultation service. implementation and prospective evaluation of a prototype. *Ann Intern Med* 1989;110:824–32.
33 Croskerry P. A universal model of diagnostic reasoning. *Acad Med* 2009;84:1022–8.
34 Hobus PP, Schmidt HG, Boshuizen HP, *et al*. Contextual factors in the activation of first diagnostic hypotheses: expert-novice differences. *Med Educ* 1987;21:471–6.
35 Ji Y, Massanari RM, Ager J, *et al*. A fuzzy logic-based computational recognition-primed decision model. *Inf Sci* 2007;177:4338–53.
36 Monteiro S, Sherbino J, Ilgen JS, *et al*. The effect of prior experience on diagnostic Reasoning: exploration of availability bias. *Diagnosis* 2020;7:265–72.
37 Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. *Med Teach* 2013;35:e1218–29.
38 Sibbald M, Sherbino J, Ilgen JS, *et al*. Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation. *Adv Health Sci Educ Theory Pract* 2019;24:427–40.
39 Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. *Acad Med* 2011;86:307–13.
40 Bordage G. Why did I miss the diagnosis? some cognitive explanations and educational implications. *Acad Med* 1999;74:S138–43.

Appendix 1: Sample cases

Sample Easy Case: Colon cancer

Chief Complaints: This patient is a 60-year-old white male who presented with a three-week history of crampy lower abdominal pain and severe anemia.

History of Present Illness: He was in his usual state of health until 2-3 weeks prior to admission when he developed crampy lower abdominal pain which was intermittent and bilateral and not clearly related to eating, bowel movements or position. On the day prior to admission, the pain worsened. He was awakened the morning of admission with pain which increased throughout the day. He presented to an urgent care facility where his hematocrit was found to be 19.3. He denied bright red blood per rectum or melena. He has had increased fatigue and denied any other symptoms, such as vomiting, hematemesis, hematuria, change in urine color, or change in bowel habits or stool. His appetite has been normal. He believed he had lost some weight but could not quantify the amount.

Past Medical History was significant for coronary artery disease, S/P bypass grafting, asthma, and eczema.

Medications included only acetaminophen. He denied medication allergies.

Social/Family History: He was a technical illustrator who has 3-4 beers each week. Family history was unremarkable.

Physical examination revealed a pale man. He was a febrile and his pulse was 78, with a respiratory rate of 18 and a blood pressure of 132/68. He did not have orthostatic hypotension. The skin had no bruises, petechiae, or jaundice. The HEENT exam was unremarkable other than pale conjunctivae. The pulmonary examination was within normal limits. The cardiac exam revealed a II/VI systolic murmur at the left upper sternal border without radiation, but no extra heart sounds or rubs. There was mild tenderness to palpation in the lower abdominal quadrants, without rebound or guarding. The liver edge was palpable 2 cm below the right costal margin and was 10 cm by percussion in the mid-clavicular line. There was no splenomegaly, nor any masses. Stool was guaiac-positive and brown. The extremities were unremarkable and no neurologic deficits were noted.

Laboratory Data

			Normal	
CBC	Hct	17.3	42-52	%
	Hgb	50	140-180	g/L
	MCV	55.4	80-100	fl
	WBC	5.2	4-10.0	X 10^9/l
	platelet count	273	200-400	X 10^9/l
Chemistries	electrolytes	within normal limits		
	creatinine	71	80-115	µmol/L
	BUN	13	8-20	mg/dl

Laboratory Data

calcium	2.20	2.15-2.55	mmol/L
phosphorus	1.1	0.8-1.6	mmol/L
protein, total	72	60-83	g/L
albumin	39	35-49	g/L
bilirubin, total	9	2-19	µmol/L
transaminases	within normal limits		
LDH	87	60-200	U/L
ALP	60	30-130	U/L
PT, PTT	normal		

Urinalysis: unremarkable

Chest X-ray: unremarkable

Sample hard case: Syphilitic meningitis

Chief Complaint: This 25-year-old woman presented with a chief complaint of headaches and tinnitus.

History of Present Illness: The patient first sought medical attention in the fall of last year. She complained of a headache that had been constant for three weeks, ringing in her ears and dizziness. The headache was severe enough that it awoke her from her sleep at times. It was not relieved by Tylenol or an over-the-counter analgesic containing aspirin, salicylamide, and caffeine. Three months previously she had been kicked in the head with a cowboy boot though did not lose consciousness. The subject of spouse abuse, she had had multiple episodes of head trauma in the past and had been struck with a baseball bat in the head two years previously. On examination she had some tenderness in the left anterior parietal skull. She underwent a head CT that showed no evidence of a subdural hematoma but questioned the presence of a nondisplaced left posterior frontal skull fracture versus a vascular groove. The patient was advised of community resources for abused women and left home for three weeks.

Two weeks later she returned because of continued tinnitus. She described it as constant though fluctuating in intensity. Voices were muffled at times. Her headache was not quite as bad. On exam she had fluid and air bubbles behind both tympanic membranes. There was slight erythema of the left TM. She was started on amoxicillin 500 mg tid, Otrivin nose drops and Dimetapp for otitis media. She took one amoxicillin capsule and developed a rash with hives, itching and a low-grade fever. She had never experienced a penicillin allergy in the past. The amoxicillin was stopped and she was started instead on Keflex.

Five days later she still complained of tinnitus and had a faint macular rash over the flexural creases of her arms. She was sent for an audiogram which showed bilateral, high frequency hearing loss. Blood work revealed a sed rate of 77 mm/hr. She was referred to this hospital for further evaluation. On admission the patient complained of continued tinnitus. She experienced occasional vertigo and flashing lights. She still had a faint rash. She denied fevers, chills, nausea, vomiting and weight loss.

Past Medical History: She had a past history of genital herpes and chlamydia infection. She had been pregnant five times, the first two ending in spontaneous abortions. The third pregnancy required a C-section at delivery because of the active herpetic lesions. From her fourth and fifth pregnancies, she delivered 36-week gestation twins by C-section 6 years ago and had a repeat C-section of a term infant 3 years ago. She denied previous drug allergies. She had a norplant for contraception.

Social History: She smokes a pack a day. She does not drink. She denied exposure to pets. She had not traveled.

Physical Exam: Her temperature was 98.7 F (37.1 C). The blood pressure was 108/64 with a pulse of 81 and respiratory rate of 18. She was well developed and well nourished. Her pupils were equal, round, reactive to light and accommodation. The fundi were normal. The TM's were normal. The oral pharynx appeared normal. She had no jugular venous distention and no adenopathy. The neck was supple. The heart sounds were normal. The lungs were clear to auscultation. The abdomen was soft and non-tender without organomegaly. She had a vertical midline scar up to the umbilicus. The pelvic exam showed no discharge, no cervical motion tenderness and no masses. She had a faint erythematous macular rash on her forearms that did not involve the palms. The neurologic exam was non-focal.

Laboratory Data

Laboratory Data

			Normal	
CBC	Hct	38	38-47	%
	Hgb	131	123-157	g/L
	WBC	8.8	4-12	X 10^9/l
	Neut	70	40-70	%
	lymph's	19	20-50	%
	mono	6	2-10	%
	eos	1	2-5	%
Chemistries	sodium	141	135-145	mmol/l
	potassium	3.3	3.5-5.0	mmol/l
	chloride	104	100-111	mmol/l
	CO2	26	24-30	mmol/l
	creatinine	106	53-106	µmol/L
	BUN	7	8-20	mg/dl
	bilirubin, total	14	0-21	µmol/L
	albumin	42	35-50	g/L
	protein, total	74	68-83	g/L
	AST (SGOT)	18	9-26	U/L
	ALT (SGPT)	13	7-30	U/L
	LDH	146	108-215	U/L
	ALP	170	39-117	U/L

Urine dip stick: negative, 3-10 WBC's, >50 epithelial cells.

Cervical cultures: negative for chlamydia and gonorrhea.

Chest X-ray: clear lung fields and normal cardiac silhouette.