



The Effectiveness of Electronic Differential Diagnoses (DDX) Generators: A Systematic Review and Meta-Analysis

Nicholas Riches, Maria Panagioti, Rahul Alam, Sudeh Cheraghi-Sohi, Stephen Campbell, Aneez Esmail, Peter Bower

Published: March 8, 2016 • <http://dx.doi.org/10.1371/journal.pone.0148991>

Abstract

Background

Diagnostic errors are costly and they can contribute to adverse patient outcomes, including avoidable deaths. Differential diagnosis (DDX) generators are electronic tools that may facilitate the diagnostic process.

Methods and Findings

We conducted a systematic review and meta-analysis to investigate the efficacy and utility of DDX generators. We undertook a comprehensive search of the literature including 16 databases from inception to May 2015 and specialist patient safety databases. We also searched the reference lists of included studies. Article screening, selection and data extraction were independently conducted by 2 reviewers. 36 articles met the eligibility criteria and the pooled accurate diagnosis retrieval rate of DDX tools was high with high heterogeneity (pooled rate = 0.70, 95% CI = 0.63 to 0.77; $I^2 = 97\%$, $p < 0.0001$). DDX generators did not demonstrate improved diagnostic retrieval compared to clinicians but small improvements were seen in the before and after studies where clinicians had the opportunity to revisit their diagnoses following DDX generator consultation. Clinical utility data generally indicated high levels of user satisfaction and significant reductions in time taken to use for newer web-based tools. Lengthy differential lists and their low relevance were areas of concern and have the potential to increase diagnostic uncertainty. Data on the number of investigations ordered and on cost-effectiveness remain inconclusive.

Conclusions

DDX generators have the potential to improve diagnostic practice among clinicians. However, the high levels of heterogeneity, the variable quality of the reported data and the minimal benefits observed for complex cases suggest caution. Further research needs to be undertaken in routine clinical settings with greater consideration of enablers and barriers which are likely to impact on DDX use before their use in routine clinical practice can be recommended.

Citation: Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. (2016) The Effectiveness of Electronic Differential Diagnoses (DDX) Generators: A Systematic Review and Meta-Analysis. PLoS ONE 11(3): e0148991. doi:10.1371/journal.pone.0148991

Editor: Robert L. Schmidt, University of Utah Health Sciences Center and ARUP Laboratories, UNITED STATES

Received: September 3, 2015; **Accepted:** January 25, 2016; **Published:** March 8, 2016

Copyright: © 2016 Riches et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: All relevant data are within the paper and its Supporting Information files.

Funding: This publication is independent research by the NIHR Greater Manchester Primary Care Patient Safety Translational Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The study was funded by the NIHR. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that (1) NR, MP, RA, SCS, SC, AE and PW have no support from any organisations for the submitted work; (2) NR, MP, RA, SCS, SC, AE and PW have no relationships with any companies that might have an interest in the submitted work in the previous 3 years; (3) their spouses, partners, or children have no financial relationships that may be relevant to the submitted work; and (4) NR, MP, RA, SCS, SC, AE and PW have no non-financial interests that may be relevant to the submitted work.

Introduction

Diagnostic error occurs when a clinician fails to make the correct diagnosis at an appropriate time or misses a diagnosis[1]. Rates of diagnostic error have been estimated at 10–15% in most areas of clinical medicine[2] and the estimated proportion of medico-legal claims against primary care doctors due to diagnostic error range between 63–72% [3, 4].

Addressing diagnostic error is complex and suggested approaches include training in diagnostic techniques for clinicians and the use of electronic diagnostic aids to augment the diagnostic abilities of doctors[5, 6]. Causes of diagnostic error are numerous but suggested solutions include training in diagnostic techniques for clinicians and the use of electronic diagnostic aids to augment the diagnostic abilities of doctors [6].

Differential diagnosis (DDX) generators are one form of electronic diagnostic aid and were developed in the 1960s[7]. These computer programmes suggest differential diagnoses based on clinical data input by users and the programmes vary in their computational methods such as utilising Bayesian probabilities and/or utilising text mining techniques. DDX programmes continue to evolve with their computational methods, particularly across medical specialities [8]. Some of the contemporary DDX generators available for generalist clinicians are capable of searching large electronic databases and are predominantly web-based providing easy access and flexibility in use while being continuously updated to reflect current evidence.

In one recent study, Bond and colleagues performed a head-to-head evaluation of four DDX generators which used clinical cases to rank them according to a set of criteria, with ISABEL and DxPlain scoring joint first in identifying the correct diagnosis[9]. One previous systematic review by Garg et. al reported improvements in practitioner performance following the use of disease and symptom-specific DDX generators[10]. In a more recent broader narrative review on the applications of information technology to the diagnostic process, El-Kareh and colleagues reported accuracy rates for DDX generators in the range of 70–95% and suggested the development of alternative metrics to measure diagnostic performance[11].

An awareness that DDX generators could help to address diagnostic error has been growing, including in a Kings Fund report into diagnostic error[12]. A scoping exercise performed prior to this review identified several studies which assessed the characteristics of DDX generators. Most studies reported ‘diagnostic accuracy’ as the primary outcome. This is not analogous to conventional definitions of test accuracy since DDX generators produce a differential diagnosis list of variable length. In this context, ‘diagnostic accuracy’ represents the proportion of searches in which the correct diagnosis appears in an output list of variable length. To emphasise this distinction we will subsequently use the term ‘accurate diagnosis retrieval’ in place of accuracy.

The literature demonstrates that accurate diagnosis retrieval alone does not predict the uptake and effectiveness of DDX generators in clinical settings. Other relevant characteristics which can have an impact on uptake and effectiveness include the specificity of the diagnostic list[9], time taken to use[13], availability and access [9], and cost-effectiveness[9].

There has been no previous systematic review of the effectiveness of DDX generators in general clinical practice. We therefore aimed to conduct a systematic review and meta-analysis to assess the clinical effectiveness of DDX generators. This was defined according to four key research questions:

1. Are DDX generators effective at retrieving accurate diagnoses?
2. Do DDX generators perform as well as clinicians?
3. Does consulting a DDX generator improve the accuracy of a clinician’s diagnostic list?
4. What are the enablers and barriers to the use of DDX generators in clinical practice?

The first three questions pertain to the efficacy of DDX generators. Their ability at retrieving accurate diagnoses was measured and compared with that of clinicians where applicable. The impact of DDX generators on the diagnostic performance was also assessed. The final question considers other factors which determine whether these tools have utility in clinical settings. Finally, we aimed to offer recommendations to researchers, policy makers and clinicians regarding the use of DDX generators in clinical practice and recommendations regarding the future research agenda in this area.

Methods

Protocol and registration

This review was conducted and reported according to Preferred Reporting Items for Systematic Reviews and Meta-analyses

Statement (PRISMA) guidelines[14] (S1 PRISMA Checklist.) and registered with PROSPERO in March 2014.

Available at: http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42014007638)

Eligibility criteria

We included any primary research study investigating the effects of DDX generators on patient care and reporting quantitative data on pre-specified outcomes. Eligible study designs included randomised controlled trials, interrupted time series analysis, cohort studies, case control studies, cross-sectional studies and before and after studies. No language restrictions were applied.

Exclusion criteria

DDX tools with a focus on a particular disease or speciality were excluded to minimise heterogeneity as well as their applicability for generalist clinicians as demonstrated by our scoping exercise.

Participants

There were two groups: the individual user of the tool and the clinical case being entered into the tool. No restrictions were made on the characteristics of individual users of DDX generators, although data on training and clinical setting were recorded to allow subgroup analysis.

Cases entered by these users could be either real clinical cases or simulated cases originating from primary or secondary care, provided they were written by clinical experts and contained diagnostic uncertainty. Cases from both paediatric and adult medical specialties were included.

Intervention

The intervention was use of a DDX generator to improve diagnostic performance. We adopted a definition of 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”*.

In order to be as comprehensive as possible, we included DDX generator tools which are no longer available.

Comparator

Exploratory work identified different comparators used to determine the effectiveness of DDX generators. The following were included in this review:

1. 'Clinical diagnosis'—used for real cases when the actual diagnosis made in practice (e.g. discharge diagnosis) is used as a proxy for gold standard diagnosis
2. 'Simulated diagnosis'—used for cases written by a panel of experts, when the consensus opinion regarding the correct diagnosis is taken as the gold standard.
3. 'Before' groups—in these studies the accuracy of clinician diagnosis is compared before and after using a DDX generator.

Outcomes

We built on previously developed criteria [15]and identified additional utility variables of relevance and extracted data on these. These are listed and defined in Table 1.

VARIABLE	DEFINITION
Accurate diagnosis retrieval	Proportion of DX tool differential lists which contain the correct diagnosis
Diagnostic detail -Comprehensiveness	Proportion of gold standard differential list which appears in the DDX differential list
Diagnostic detail—Number of diagnoses	Average number of diagnoses generated during each use of a DDX tool
Diagnostic detail—Relevance	Clinical appropriateness of DDX tool output
Diagnostic detail—Diagnostic list	Impact of using the DDX tool on the diagnostic list made by clinicians (e.g. adding / removing diagnoses)
Usage data—Time to use	Average time taken for a user to operate a DDX tool for a given case
Usage data—Frequency of use	How often DDX tools are used in a clinical setting
Usage data—Satisfaction	User satisfaction with DDX tools (can relate to patient management or educational benefits) / and/or relate to the likelihood of DDX tools being accepted and utilised by clinicians
Moderators of outcomes—Case difficulty	Impact of case complexity on DDX tool outcomes
Moderators of outcomes—Clinical experience	Impact of user's clinical experience on DDX tool outcomes
Outcomes—Investigation	Impact of using DDX tools on the ordering of diagnostic investigations
Outcomes—Cost-effectiveness	Economic impact of using DDX tools in a clinical setting

doi:10.1371/journal.pone.0148991.t001

Table 1. Description of efficacy and utility variables.
<http://dx.doi.org/10.1371/journal.pone.0148991.t001>

Information sources and searches

The following databases were searched from inception to November 2013 and updated in June 2015: Ovid MEDLINE(R), Embase, CINAHL, Cochrane Database of Systematic Reviews, Cochrane Central Register of Controlled Trials, Cochrane Methodology Register, ACP Journal Club, Database of Abstracts of Reviews of Effects, Health Technology Assessment, NHS Economic Evaluation Database, AMED (Allied and Complementary Medicine), CAB Abstracts, Global Health, Health and Psychosocial Instruments, Health Management Information Consortium and PsycINFO. A combination of MeSH terms and text words were used describing medical diagnosis, including electronic diagnosis and the names of specific DDX generators which had been identified from exploratory work.

In addition to the websites identified by the search, hand-searches of the websites of the National Patient Safety Agency (NPSA) and the Agency for Healthcare Research and Quality (AHRQ) were undertaken. The reference lists of included articles were also screened for eligible papers and we conducted Scopus searches for all articles citing the included studies.

The complete search strategy is available in S1 File. The search was not restricted by date, language or country of publication.

Study selection

A two-stage data selection process was followed. 1) Titles and abstracts were screened and 2) full-texts of the eligible titles and abstracts were retrieved and reviewed against the eligibility criteria. Both stages were independently completed by two reviewers and any disagreements were resolved in group meetings until consensus was reached. High inter-rater reliability was achieved: Cohen's[16] unweighted κ coefficient = 0.88 and 0.91 for title/abstract and full-text screening, respectively.

Data extraction

A standardised data extraction form was developed and piloted. Studies meeting the inclusion criteria were then double-extracted by the review team. The first author extracted data from all of the included studies to ensure consistency. Data were extracted and cross-checked by pairs of reviewers using the Microsoft Excel data extraction form. Disagreements were resolved by discussion in group meetings.

Participants—Extracted data on user characteristics included the total number of users in each study, clinical background, clinical grade and whether they had been trained to use the DDX generator. Case characteristics included the type of case (e.g. real vs. simulated), clinical specialty, setting (e.g. primary or secondary care), the number of unique clinical cases included in each study and the combined number of clinical cases in each study (since one case could be used by numerous users in a single study).

Intervention—We extracted data related to the type of DDX generator used, whether a complete or abbreviated list of differential diagnoses was analysed and whether it was used in real-time (e.g. prospectively or retrospectively).

We extracted data on the type of comparator used (e.g. gold standard or a 'before' group) as well as the type of gold standard diagnosis being used (e.g. published case report diagnosis or discharge diagnosis).

Outcomes—For each study we extracted any available data for the outcomes listed in Table 1.

Risk of bias in individual studies

An adapted version of the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) was used to assess the methodological quality of the studies included in the review (S1 File)[17]. QUADAS-2 is specifically designed for the quality appraisal of diagnostic studies. It usually consists of seven domains; four relate to risk of bias and three relate to applicability. The scoping review highlighted several studies which had been funded by the software manufacturers which we identified as a potential source of bias. We therefore chose to incorporate an additional domain of "commercial funding" within our adapted version of QUADAS-2.

Suggested 'signalling questions' for each domain were tailored to this study by the authors (S1 File). Each domain was then scored as high, low or unclear. Studies were not excluded from the review on the basis of quality, but the results of the methodological quality assessment were used in the interpretation of the results.

Data synthesis and analysis

Accurate diagnosis retrieval was the primary outcome of this review. The majority of the studies ($n = 21$ out of 33) reported this as a crude proportion without incorporating control groups. Rates of accurate diagnosis retrieval of the DDX generators across the studies were extracted and pooled using the single-group mode of the Comprehensive Meta-Analysis (CMA) version 2.23[18]. Subgroup analyses were performed to examine whether different types of DDX generators were associated with different levels of accurate diagnosis retrieval. We also conducted sensitivity analyses to examine whether the results altered when studies with high methodological quality ratings (based on QUADAS-2) and when DDX generators that are currently commercially available were retained in the analyses.

Two-group meta-analyses were undertaken for 7 studies which compared the accurate diagnosis retrieval rates of the DDX generators with alternative diagnostic approaches (i.e., clinical diagnoses by doctors and students) and 5 studies which examined accurate diagnosis retrieval rates before and after the use of DDX generators (before and after studies; $n = 5$) [19–23]. Accurate diagnosis retrieval data from these two groups of studies were extracted and converted into a common effect size (Standardised Mean Difference; SMD) and pooled in CMA. A positive SMD indicated that DDX generators were associated with higher levels of accurate diagnosis retrieval whilst a negative SMD indicated that DDX generators were associated with lower levels of accurate diagnosis retrieval [24]. In keeping with established cut-off points of effect, effect sizes of 0.7 and higher were categorised as large; effect sizes of 0.30 to 0.60 as moderate, and effect sizes 0.2 and lower as small [25]. A random effects model was used throughout to control for between-study heterogeneity. STATA software (version 13) was used to create the forest plots.

The Cochran's Q statistic [26] and the Higgin's I^2 [26] statistic were used to assess between-study heterogeneity. The Q statistic provides an estimate of whether differences between meta-analysed studies are greater than would be expected by chance. Statistically significant results indicate the presence of heterogeneity. The I^2 statistic provides a quantitative measure of the degree of between study differences caused by factors other than sampling error. Higher I^2 values represent greater heterogeneity [26]. Publication bias was examined using a test of funnel plot asymmetry (Egger's test) [27] and Rosenthal's fail safe N (FSN) [28]. Egger's test reveals whether or not the funnel plot is symmetric and indicates the existence/absence of a significant publication bias, and the FSN provides an estimate of the number of studies with statistically non-significant results are needed for a meta-analysed finding to become statistically non-significant.

A considerably lower number of studies reported data on the utility of DDX generators compared to studies reporting data on accurate diagnosis retrieval. A wide range of outcomes were described as "utility" outcomes, however due to the low number of studies and the inconsistent reporting of data, utility outcomes were precluded from a meta-analyses. Hence, a narrative synthesis was undertaken for utility outcomes.

Results

The PRISMA flowchart demonstrates the screening and selection process for the review and is outlined in Fig 1. The search generated 9299 references and following initial screening, 92 articles were short-listed for full text screening. Of these, 36 articles met the eligibility criteria [9, 13, 15, 19–23, 29–56]. Two articles [46, 55] included 2 discrete studies on the same DDX generator and these were considered as separate studies for the purposes of this review leading to a total of 38 eligible studies. Additionally, 6 articles reported data on more than one DDX generator and these were also considered as separate studies [9, 13, 15, 19, 39, 51]. This resulted in a total of 48 independent DDX generator comparisons reporting either diagnostic accuracy; clinical utility data, or a combination.

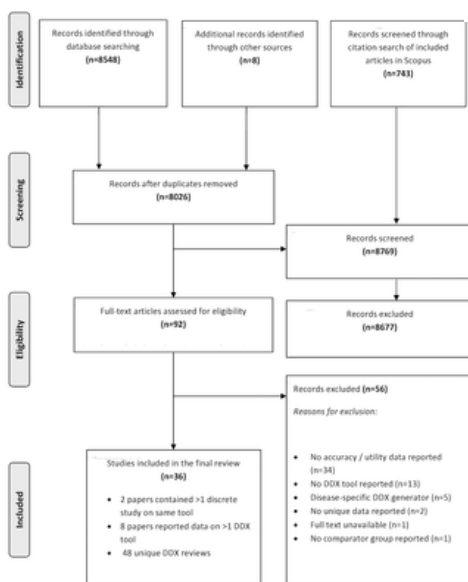


Fig 1. PRISMA flowchart.

<http://dx.doi.org/10.1371/journal.pone.0148991.g001>

Study characteristics

Complete study characteristics are detailed in Table 2. In total, 36 articles provided data on 11 different DDX tools, of which 6 are known to be no longer commercially available.

STUDY ID	INDEX TEST	COUNTRY	DESIGN	SETTING	SPECIALTY	CASE TYPE	CASES ANALYSED	TIME PERIOD	INVESTIGATOR BACKGROUND	DATA ENTRY
Adair 1988 (1)	QMR	USA	RCT	Primary care	General practice	Real	1602	1982	Other	Retrospective
Adair 1988 (2)	QMR	USA	RCT	Hospital	General medicine	Simulated	48	1982	Other	Retrospective
Baichwal 1988 (1)	QMR	Canada	RCT	Hospital	General medicine	Simulated	9	96	Chiropractic	Prospective
Baichwal 1988 (2)	QMR	Canada	RCT	Hospital	General medicine	Real	20	20	Chiropractic	Prospective
Barnes 1988 (1)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	10	80	Chiropractic	Retrospective
Barnes 1988 (2)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (3)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (4)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (5)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (6)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (7)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (8)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (9)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (10)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (11)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (12)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (13)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (14)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (15)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (16)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (17)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (18)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (19)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (20)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (21)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (22)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (23)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (24)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (25)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (26)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (27)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (28)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (29)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (30)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (31)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (32)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (33)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (34)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (35)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (36)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (37)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (38)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (39)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (40)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective
Barnes 1988 (41)	ISABEL	USA	DAS	Hospital	Pediatrics	Real	63	63	Chiropractic	Retrospective

Table 2. Study characteristics.
<http://dx.doi.org/10.1371/journal.pone.0148991.t002>

The majority of included studies were diagnostic accuracy studies (n = 33 discrete DDX tool investigations) comparing diagnostic accuracy using a DDX generator with a pre-determined gold standard diagnosis[9, 13, 15, 19–22, 30–34, 36–41, 43–47, 49–56]. Five articles reported no accuracy data[23, 29, 35, 42, 48] and 29 studies from 28 articles reported at least one component of utility data[13, 15, 19–23, 29–48, 55] (one article comprised of 2 separate studies containing utility data [46]).

5 articles contained control groups, 3 of which were observational studies[35, 42, 48] and 2 were randomised controlled trials[29, 31]. Summary characteristics of all the included studies are presented in Table 2.

Risk of bias

Quality scores were variable with more recent studies demonstrating higher quality[9, 23, 29, 35, 37, 49]. The risk of bias for individual studies is listed in S2 File.

When the risk of bias was summarised across studies (Fig 2), a particularly high risk of bias was demonstrated in relation to case selection. This was typically due to a lack of randomization or inappropriate exclusions. Applicability was generally good for case selection and the reference standard. However, this was not the case for the index test given that most studies were not performed in real-time clinical settings. Only three studies received low risk of bias scores across seven or more criteria[9, 35, 57]. 11 studies received low risk of bias scores across 5 criteria[20, 29, 31, 36, 38, 39, 41, 47–49] and the remaining studies were deemed to have a high risk of bias [13, 15, 19, 21–23, 30, 33, 34, 40, 42–46, 50–56, 58].

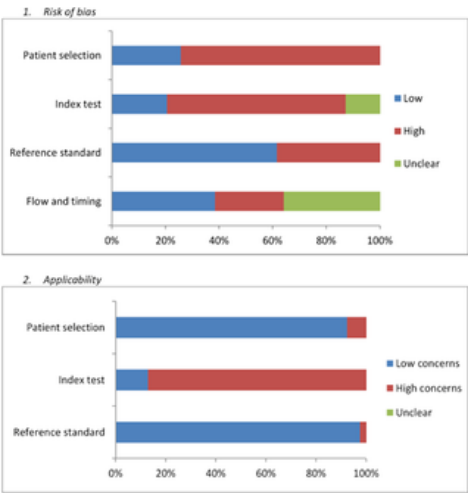


Fig 2. Risk of Bias summary table.
<http://dx.doi.org/10.1371/journal.pone.0148991.g002>

Synthesis of results

Data across the vast majority of the studies reporting accurate diagnosis retrieval data (28 of 33) were included in single-group meta-analysis to examine the first research question. The majority of these studies (21 of 28) did not include a comparator; rather the correct diagnoses were confirmed a priori by expert clinician diagnoses.

Only 7 of the 33 studies compared the accurate diagnosis retrieval of DDX generators in assigning the correct diagnoses against other diagnostic methods (e.g. clinician diagnosis) in addition to the pre-assigned expert clinical diagnoses [30–32, 38, 43, 47, 56]. These 7 studies were initially included in single-group meta-analysis (n = 28) and then were further analysed in 2-group meta-analysis to examine the second research question. The remaining 5 of the 33 studies reported accurate diagnosis retrieval data before and after the use of DDX generators. A separate two-group meta-analysis was undertaken for these 5 studies to examine the third research question.

In relation to the fourth research question, data on 11 different utility outcomes representing enablers or barriers to the use of DDX generators in clinical practice were identified but these were reported inconsistently across the studies preventing a meta-analysis, therefore a narrative synthesis of these outcomes was undertaken. The DDX tool specific data on utility outcomes are presented in S2 File.

Are DDX generators effective at retrieving accurate diagnoses?

Five of the 28 studies provided accurate diagnosis retrieval data on multiple DDX generators [9, 13, 19, 39, 51] resulting in a total of 38 independent samples included in the meta-analysis (see forest plot in Fig 3).

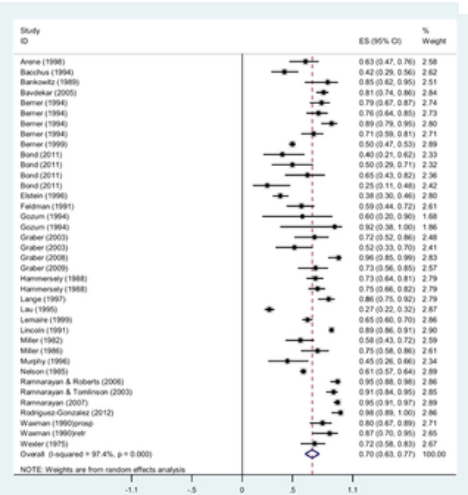


Fig 3. Single-group meta-analysis: Forest plot of the total accurate diagnosis retrieval rates of DDX generators.

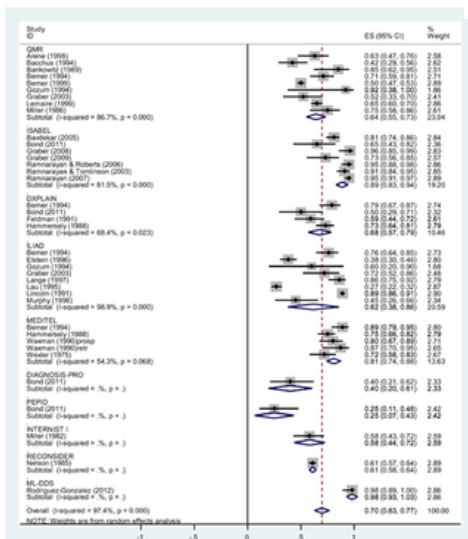
Heterogeneity chi-squared = 1414.31 (d.f. = 37), $p < 0.0001$. Note: Random effects model used. 95% CI = 95% confidence intervals; ES = rates.

<http://dx.doi.org/10.1371/journal.pone.0148991.g003>

The pooled accurate diagnosis retrieval rate of DDX tools was high but the heterogeneity was also high (pooled rate = 0.70, 95% CI = 0.63 to 0.77; $I^2 = 97\%$, $p < 0.0001$) (see Fig 3). The individual rates ranged widely from 0.25 in a study examining the accurate diagnosis retrieval rate of the PEPID generator[9] to 0.98 in a study examining the accurate diagnosis retrieval rate of the ML-DDS generator based on only 5 cases[56]. As shown in Fig 3, a total of 11 studies reported accurate diagnosis retrieval rates lower than 60 percentage points and 13 studies reported accurate diagnosis retrieval rates higher than 0.80.

Subgroup analysis.

A subgroup analysis was conducted to examine whether different types of DDX generators were associated with different levels of accurate diagnosis retrieval (see forest plot in Fig 4). ISABEL[9, 22, 23, 34, 37, 38, 45, 49, 55] was associated with the highest rates of accurate diagnosis retrieval compared to all other types of DDX tools, but heterogeneity was high (pooled rate = 0.89, 95% CI = 0.83 to 0.94; $I^2 = 82\%$, $p < 0.0001$). MEDITEL[39, 46, 47] was also associated with high rates of accurate diagnosis retrieval with moderate heterogeneity (pooled rate = 0.81, 95% CI = 0.74 to 0.88; $I^2 = 54\%$, $p = 0.07$). Moderate rates of accurate diagnosis retrieval were observed for DXPLAIN[9, 39, 50] (pooled rate = 0.68, 95% CI = 0.57 to 0.79; $I^2 = 68\%$, $p = 0.02$) ILIAD[13, 19–21, 36, 40, 51–53] (pooled rate = 0.62, 95% CI = 0.38 to 0.86; $I^2 = 99\%$, $p < 0.0001$) and QMR[13, 19, 30–33, 41, 43, 51] (pooled rate = 0.64, 95% CI = 0.55 to 0.73; $I^2 = 87\%$, $p < 0.0001$) but again heterogeneity was high. In one article, accurate diagnosis retrieval rates for DIAGNOSIS PRO and PEPID were reported and demonstrated low rates of accurate diagnosis retrieval [9]. Three other types of diagnostic tools (INTERNIST-1, ML DDSS, RECONSIDER) were reported in 3 articles[44, 54, 56] but again these rates were based on the results of a single study and these tools are not commercially available.

**Fig 4. Single-group meta-analysis: Forest plot of the accurate diagnosis retrieval rates by subgroups of DDX generators.**

Chi-squared values: DXPLAIN = 9.64 (d.f. = 3), $p = 0.023$, ILIAD = 634.49 (d.f. = 7), $p < 0.0001$, ISABEL = 33.61 (d.f. = 6), $p < 0.0001$, MEDITEL = 8.73 (d.f. = 4), $p = 0.068$, QMR = 60.24 (d.f. = 8), $p < 0.0001$. Note: Random effects model used. 95% CI = 95% confidence intervals; ES = rates.

<http://dx.doi.org/10.1371/journal.pone.0148991.g004>

Sensitivity analyses.

From the pooled single-group meta-analyses we removed studies which were assigned high risk of bias ratings across 4 or more criteria. Eighteen studies including 28 independent samples were retained in the meta-analysis [9, 13, 15, 21, 32, 33, 36–41, 44, 46, 47, 49, 51]. The pooled effect size was slightly lower compared to the pooled effect size obtained in the overall analysis (pooled rate = 0.68, 95% CI = 0.61 to 0.74, $I^2 = 93.1\%$, $p < 0.0001$) (see forest plot in Fig 5).

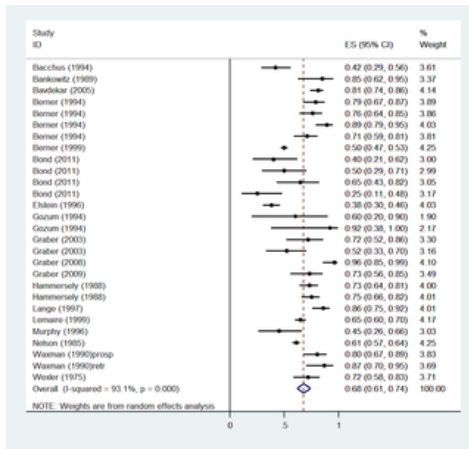


Fig 5. Sensitivity analysis: Forest plot of the rates of accurate diagnosis retrieval of DDX generators across studies with stronger methodological quality ratings.

Heterogeneity chi-squared = 267.70 (d.f. = 13), $p = 0.000$. Note: Random effects model used. 95% CI = 95% confidence intervals; ES = rates.

<http://dx.doi.org/10.1371/journal.pone.0148991.g005>

An additional sensitivity analysis was performed in which only studies reporting commercially available DDX generators were retained in the analysis (10 studies reporting 13 independent samples) [9, 15, 22, 37–39, 45, 49, 50, 55]. A slightly increased pooled effect size was found for the commercially available DDX generators (pooled rate = 0.74, 95% CI = 0.66 to 0.82, $I^2 = 92\%$, $p < 0.0001$) compared to the pooled effect size of the main analysis (see Fig 6). This result however was largely affected by the poor accurate diagnosis retrieval rates of PEPID and Diagnosis PRO which were only reported by a single study. A substantially higher pooled rate was obtained when only the 2 commonest commercially available DDX generators were retained in the analysis (pooled rate = 0.81, 95% CI = 0.74 to 0.88, $I^2 = 89\%$, $p < 0.0001$ (data not shown)).

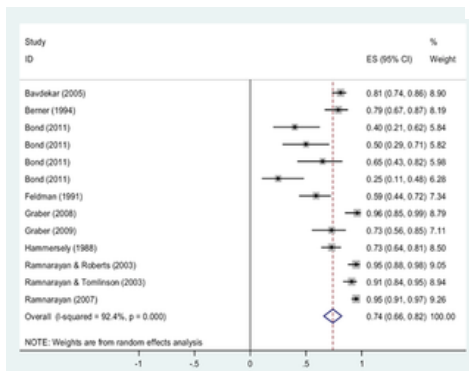


Fig 6. Sensitivity analysis: Forest plot of the rates of accurate diagnosis retrieval of DDX generators across studies testing commercially available DDX generators.

Heterogeneity chi-squared = 158.80 (d.f. = 12), $p < 0.0001$. Note: Random effects model used. 95% CI = 95% confidence intervals; ES = rates.

<http://dx.doi.org/10.1371/journal.pone.0148991.g006>

Publication Bias.

No funnel plot asymmetry (see Fig 7) was identified and Egger test was non-significant suggesting that no publication bias is present (regression intercept = 0.42, SE = 0.21, $p = 0.053$). Moreover, the FSN test indicated that as many as 1056 studies would be needed to nullify the significant effects obtained in the single-group meta-analysis.

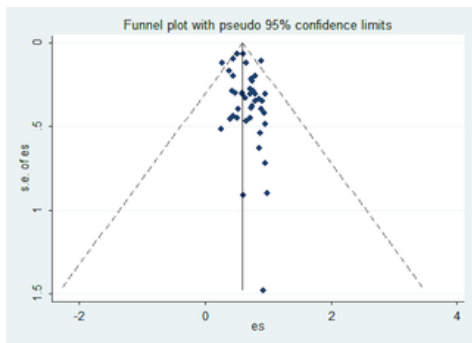


Fig 7. Funnel plot for studies examining the accurate diagnosis retrieval rates of DDX generators.

Note: ES = rates, SE of SE = standard error of rates.

<http://dx.doi.org/10.1371/journal.pone.0148991.g007>

Do DDX generators perform as well as clinicians?

The pooled standardised mean difference (SMD) of the 7 studies [30–32, 38, 43, 47, 56] which compared the efficiency of DDX tools with a comparator (clinical diagnoses by doctors ($n = 6$) and students ($n = 1$)), indicated that the use of DDX tools was associated with small, non-significant increases in accurate diagnosis retrieval compared to other ways of assigning diagnoses but the heterogeneity was high (SMD = 0.12, 95% CI = -0.30 to 0.53, $I^2 = 72\%$, $p < 0.0001$; Fig 8). ISABEL was associated with the highest accurate diagnosis retrieval rates compared to all other DDX generators[38].

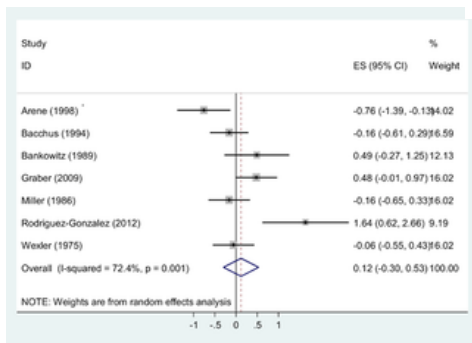


Fig 8. Two-groups meta-analysis: Forest plot of the accurate diagnosis retrieval rates of DDX generators compared to other types of clinical diagnoses.

Heterogeneity chi-squared = 21.70 (d.f. = 6), $p = 0.001$. Note: Random effects model used. 95% CI = 95% confidence intervals; ES = Standardised mean difference.

<http://dx.doi.org/10.1371/journal.pone.0148991.g008>

Does clinician use of a DDX generator after initial diagnosis lead to more accurate diagnoses?

Five studies[19–22, 34] including six independent samples examined differences in accurate diagnosis retrieval rates before and after the use of DDX tools. The pooled effects of the DDX tools indicated small, significant improvements in accurate diagnosis retrieval but the clinical significance of these minor benefits is uncertain (SMD = 0.15, 95% CI = 0.09 to 0.21; Fig 9).

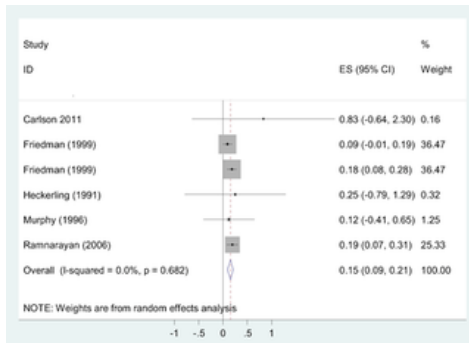


Fig 9. Two-groups meta-analysis: Forest plot of the accurate diagnosis retrieval rates of DDX generators in before-after studies.

Heterogeneity chi-squared = 2.29 (d.f. = 4) p = 0.682. Note: Random effects model used. 95% CI = 95% confidence intervals; ES = rates.

<http://dx.doi.org/10.1371/journal.pone.0148991.g009>

What are the enablers and barriers to the use of DDX generators in clinical practice?

The complete utility data is contained in S3 File, grouped by DDX generator tool. This section summarizes this evidence according to the previously identified utility variables.

Diagnostic detail.

The *relevance* of a DDX list was primarily reflected by the position of the correct diagnosis in the DDX list although conventions for truncating and displaying differential lists varied across systems (see S2 File). In two studies, the correct diagnosis appeared 1st in the differential list in 28%[44] and 23%[41] of cases and in 4 other studies the correct diagnosis appeared in the top 10 list for 78% [45], 68%[44], 51%[13], and 44%[13] of cases. The mean ranking of the correct diagnosis was reported in 4 studies at 2.3[20], 9th[39], 10th[46] and 10.7[39]. Additionally, one study comparing four tools framed relevance on the proportion of DDX tool-generated diagnoses felt to be appropriate by experts and reported values of 46%[15], 26%[15], 23%[15] and 21%[15]. Five others reported inconclusive and heterogeneous forms of individual study relevance scores[15, 19, 22, 33, 45].

Six studies[15, 20, 33, 41, 43, 55] reported on the *comprehensiveness* of the differential diagnosis output list across five tools and the scores varied according to the DDX generator: DxPlain = 38% (33–44%)[15]; ILIAD = 27% (22%–32%)[15] and 56%[20]; ISABEL = 87%[55]; MEDITEL = 39% (32–46%)[15]; and QMR = 30% (25–35%)[15], 22%[33] and 89%[41].

Nine studies[15, 19, 21–23, 32, 36, 42, 43] reported the impact of DDX generators on the content of user's *diagnostic lists*. In 6 of these studies, the use of DDX tools were associated with an increased length of diagnostic list,[15, 19, 21–23, 32] one indicated a decline in the quality of the diagnostic list[42] and one indicated no impact[43]. In two studies it was reported that the correct diagnosis was prompted by the DDX generator but then ignored by clinicians[23, 36] while another study found cases where the correct diagnosis was removed following DDX generator-consultation[19] (6.3% of cases when using ILIAD and 5.8% when using QMR). Ramnarayan also found that using ISABEL led to a significant reduction in the number of incorrect diagnoses[23].

Finally, there were seven studies where data was provided on the *number of diagnoses* generated by DDX generators[30, 32, 38, 39, 41, 47, 55]. The mean number of diagnoses generated by historical tools such as MEDITEL (40.6[47] and 46.5[39]) and QMR (48.5[30]) were greater than the most recent tool, ISABEL (13[55] and 30[38])

Usage data.

Of the 36 studies, only 3 utilised the DDX tools in real time and one of the 36 did not report when the tool was utilised. Two studies utilised the tools in primary care, 26 in hospitals, one in an academic setting and 7 studies were unclear.

There were six comparisons from five studies which reported data on the time taken to use DDX generators[22, 23, 32, 37, 39]. Historical tools ranged from 22 minutes for MEDITEL[39], 30 minutes for DxPLAIN[39] and up to 240 minutes per case episode for QMR[32]. By contrast, ISABEL was found to take between 98 seconds and six minutes on average per case in two studies[22, 23] and less than a minute in another[37].

Two studies reported data on the *frequency of use* of DDX generators and they both related to the use of the ISABEL tool. In a study by Graber, 56% of students randomized to use ISABEL to solve a clinical problem actually made use of it[38]. In a study where clinicians were given open access to ISABEL, only 7.9% reported using it more than weekly in one instance and in another study in the same paper, 54% of clinicians completed all 12 allocated cases using ISABEL[22].

Finally, seven studies reported on user *satisfaction rates* with DDX generators and they generally reported satisfaction rates to be high,[23, 30–32, 34, 43, 48] although subject to reporting bias. Some users suggested that it may be most beneficial as an educational aid for teaching diagnostic skills in a simulated environment[23, 34].

Moderators of outcomes.

Six studies considered the *clinical experience* of clinicians as a moderator of outcomes[21–23, 34, 36, 40]. When the impact of DDX generators on accurate diagnosis retrieval was compared between different clinician grades; generally the inexperienced users such as medical students[21, 22, 34, 36] and Senior House Officers[22, 23] benefited the most. The way in which users operate the tool also appeared to vary with experience with inexperienced users inputting more data into the tool[36]. Medical students were also significantly more likely to add diagnoses to a pre-existing differential list compared with more senior doctors[21, 22].

Four studies reported on the impact of *case difficulty*[22, 30, 33, 44] and found that the accurate diagnosis retrieval rates of DDX generators are lower for complex cases than easier ones[22, 30, 33] and are also lower in cases where there are multiple diagnoses[44].

Outcomes.

In addition to the impact of DDX generators on the diagnostic process itself, two other important outcomes which were reported related to the impact of using DDX generators on *investigation requests*, reported by four studies[22, 23, 31, 47], and on overall *cost-effectiveness*, reported by two studies[29, 35]. Ramnarayan et al. found that at least one significant investigation was added to the management plan following ISABEL consultation in 9.3% of cases[22]. In a separate investigation published in the same paper he found that the average number of tests ordered tended to increase following use of ISABEL[22]. Another study found that the quality of investigations arranged following DDX consultation did not increase[31] and one reported a decrease in the number of unnecessary investigations being ordered [47].

Apkon et al. found a significant increase in laboratory testing and total resource consumption for the group randomized to receive PKC[29]. By contrast Elkin found that total hospital admission costs were significantly reduced in a cohort of patients whose clinicians were using DxPLAIN, versus normal care, although there was no difference in total length of stay[35].

Discussion

Summary of evidence

Overall, this systematic review provides evidence that DDX generators have the potential to retrieve accurate diagnoses, albeit occasionally via lengthy lists. Consistent with previous reports the pooled accurate diagnosis retrieval rate of the DDX generators was 70%[11]. Commercially available DDX generators were associated with a higher accurate diagnosis retrieval rate (pooled rate = 81%) with some of the newer tools exhibiting the highest accurate diagnosis retrieval rates when compared to a gold standard. A small number of studies which compared the performance of DDX generators with the performance of primarily clinicians suggested that DDX generators were as likely as clinicians to include the correct diagnosis. Moreover, preliminary evidence from studies assessing accurate diagnosis retrieval by clinicians before and after the use of DDX generators indicated small but significant improvements in the ability of clinicians to assign the right diagnosis following the use of DDX generators.

The reporting of utility outcomes was variable. Breaking down utility data by DDX tool was not possible due to the low numbers of outcomes reported for some tools. Therefore the outcomes were combined across DDX tools for meaningful interpretation. This introduces error given the differences in DDX generators included which represent various iterations of tools over several years. Nevertheless, the majority of reported trends are consistent, albeit the heterogeneity prohibits calculating effect sizes with any certainty.

Although these findings are encouraging, they should be interpreted in the light of three key caveats, namely, that accurate diagnosis retrieval was assessed using a simplistic and unconventional manner (simple crude rates), the poor methodological quality of the included studies (including conflict of interest in some studies funded by DDX generator software manufacturers) and the high between-study heterogeneity. Key sources of the high levels of clinical heterogeneity were variations in the participants, cases and outcomes between studies. Studies included academics, students and clinicians of different grades and a mixture of real and simulated cases of varying complexities. Methodological heterogeneity was illustrated by the large variations in study designs and the high risk of bias demonstrated by some studies.

Unlike previous narrative reviews [11, 59] this is the first review which has been conducted using systematic methods including a meta-analysis which formally highlights a range of heterogeneity sources. We strongly recommend future high quality research in this area because no firm conclusions can be reached about the efficacy and utility of DDX generators based on the currently available evidence.

Diagnostic detail.

Most DDX generators produce extensive lists. As a result the likelihood of having the 'correct' diagnosis listed increases but the value of the differential list to a clinician may decrease. Moreover, studies have demonstrated that it is possible for users to miss the correct diagnosis in a DDX output list and this effect is likely to be higher the longer the list is[23]. Such long lists may increase uncertainty in clinicians, which could prevent further uptake. Additionally, the relevance and comprehensiveness of DDX generator lists were generally low. This makes the task of identifying a correct or helpful diagnosis harder for busy clinicians.

Usage data.

A reduction in the time taken to use the DDX generators was seen with newer tools such as Isabel and this may increase their acceptability in routine clinical practice. It is likely that advances in computer software designs and processing speeds in recent years have contributed to this. However, when DDX generator use was optional, the usage rates in two recent studies were generally low[22, 38]. Although there were a limited number of studies, it may suggest an element of unfamiliarity and scepticism from clinicians. Newer web-based interfaces of some tools such as Isabel may provide additional accessibility and improve future use as well their ability to utilise a wider set of remote databases leading to higher accurate diagnosis retrieval rates. Despite this, evidence of satisfaction levels in studies which ensured exposure to the tools, indicated a high level of satisfaction and this could be down to the novel experience as well as the potential benefits of using DDX generators[23, 30–32, 34, 43, 48].

Moderators of outcomes.

The data suggested that junior members of the clinical team (e.g. medical students) inputted more data and were more likely to benefit from use of these tools. This is unsurprising given their lower levels of experience and this may have an important role to play in education and training in diagnostic techniques. Additionally, the relevance and accurate diagnosis retrieval rates of DDX generator outputs fell in the context of complex cases[22, 30, 33, 44], which is presumably when a DDX generator is most likely to be needed.

Outcomes.

The evidence in relation to the number of investigations ordered and cost-effectiveness is limited by the small number of studies reporting relevant data. There was a trend towards use of DDX generators increasing the numbers of investigations requested[22, 29], but this was not seen consistently[32, 47]. Data on cost-effectiveness were also inconsistent[29, 35].

Research and policy implications

This review demonstrates that DDX generators, particularly more contemporary versions have high accurate diagnosis retrieval rates when used in an experimental setting. The significance of this finding must be interpreted with caution however. As stated, the length of a DDX generator diagnostic list is a key predictor of accurate diagnosis retrieval. Long diagnostic lists are less specific and hence problematic for clinicians using these tools in a busy clinical setting. Moreover, the majority of DDX tools had no ability to rank the order of diagnoses and such a function is likely to have added value. Where rank/ordering of diagnoses were available, they were often limited to a small number of tools and varied significantly with their range.

Studies exploring the efficacy and utility of tools in prospective clinical settings are limited in number and quality. Recommendations are limited by the low quality of most included studies, their varying research designs, methodologies and heterogeneous outcomes.

Research should be conducted in a prospective, generalist clinical setting, ideally with cases stratified according to their complexity and users stratified according to their clinical experience. Given the heterogeneity and scarcity of high quality evidence we recommend a standardised and progressive approach (similar to that used by others such as the MRC complex interventions framework[60]) in developing DDX tools for use by clinicians in their routine clinical workflow. Although unlikely to be a priority for commercial entities, work must begin with understanding the barriers, facilitators and preferences to utilising DDX tools in routine clinical practice by professionals and they must have an understanding of patient perspectives. Such work should be followed by small scale exploratory studies such as controlled before-and-after studies exploring the impact on all relevant outcomes from diagnostic retrieval rates to the impact on ordering of tests/investigations and cost-effectiveness. We would recommend such studies receive funding independent of the software manufacturer to ensure scientific rigour.

Given the variance in the literature, where programmes are unable to rank diagnoses in order, at the least, developers should consider offering probabilities based on patient presentation, patient characteristics linked to electronic health records and potentially patient demographics. Where outcomes such as ranking are absent, we suggest standardisation of an “accurate retrieval” to be the correct diagnosis listed amongst the top-5 diagnoses produced. Following this, rigorous formal trials can be used to assess causal impact followed by pragmatic large-scale cost effectiveness studies with long enough follow-up periods to measure impact on patient safety, outcomes and costs.

At present, there is insufficient evidence to recommend the uptake of DDX tools in clinical settings. However, the data suggests a potential role for these tools in teaching diagnostic skills in a simulated setting.

Strengths and Limitations

This systematic review had several strengths. The study eligibility criteria were broad to allow for a comprehensive overview of published data in this specialist area for generalist clinicians. The decision to include DDX tools that are no longer commercially available was helpful since much of the evidence relates to these, and much of the data relating to these is consistent with that found in more recent studies. Searches involved screening multiple ($n = 16$) databases supplemented by hand searches of the reference lists of studies included in the review. There is evidence that the non-inclusion of grey literature findings is associated with larger intervention effects[61]. To reduce this possibility in the case of DDX generators, grey literature was included in this review. Screening and data extraction were completed by two independent researchers and demonstrated very high levels of

agreement. Exploratory work prior to the review allowed for a categorization of 'utility' concepts which matched the extracted data well. This review focused on both efficacy and utility because these are supplementary elements which determine the overall effectiveness of DDX tools.

Due to the nature of these tools and the low reported quality of the included studies there are limitations. Firstly, comparing the retrieval rates of different systems in different contexts using different cases is suboptimal resulting in our recommendation for future standardised work. The studies exhibited high heterogeneity and the main meta-analysis was based on pooling crude accurate diagnosis retrieval rates from studies lacking control groups. Although subgroup and sensitivity analyses were performed to explore key sources of heterogeneity (type of DDX generators, current availability and methodological quality), some subgroup analyses were based on a very small number of studies which do not allow the formation of robust conclusions. In addition, the high heterogeneity demonstrated by the meta-analysis could be attributed to other major variations such as study design and user and case characteristics which have not been accounted for in the analyses.

We pursued meta-analysis in this review because it facilitates the comparison of the results across studies, the examination of the consistency of effects, and the exploration of key variables that might account for inconsistencies[62]. A narrative synthesis does not allow such useful manipulations. In the light of the large between-study variations however, the findings of the meta-analysis should be interpreted with caution.

Some tools such as ILIAD[13, 15, 19–21, 36, 40, 42, 48, 51–53] and ISABEL[9, 22, 34, 37, 38, 45, 49, 55] was examined more often than others; some as low as once[9, 29, 44, 54, 56]. This made direct comparison between tools problematic. Furthermore the risk of bias assessment highlighted that some studies were partly funded by the DDX generator software manufacturers themselves, which is likely to introduce funder and publication bias. The significant risk of bias posed by this element requires careful consideration when interpreting the findings. Linked to this bias were concerns of the inconsistent and poor reporting of the data entry personnel, primarily whether they were funder employees, researchers or clinicians and whether they entered the data in real time or post-hoc.

Moreover, for the majority of studies, the accuracy of the systems was determined by measuring accurate diagnosis retrieval rates. In the absence of restrictions on the length of a diagnostic list, the clinical relevance of this concept is questionable. These retrieval rates were also based on the premise that gold standard clinician diagnoses are always correct and the review's findings are heavily reliant on the internal validity of the primary studies. These issues are likely to impact on both the efficacy and utility of DDX generators. Whilst some of the older tools took longer to use, some of the newer tools demonstrated significantly lower time-taken to use. It is likely that improvements in information technology as well as advances between the interfaces of different health systems has contributed to the reduction in time and this may have important implications for use in routine clinical practice.

Another limitation of the study is that, although we went to extreme lengths to capture all relevant studies, it may be possible that DDX programmes exist on the market that have not been subject to peer-reviewed publications and as such have not been included in this review. Finally, a key barrier to the external validity of these results is that DDX generators were applied retrospectively across the majority of the studies. The tools must be used prospectively in order to reduce diagnostic error as part of a busy workload.

Conclusions

Our findings demonstrate that DDX generators have the potential to improve diagnostic practice and thereby reduce diagnostic error, but there is currently insufficient evidence from the existing literature to recommend their routine use by clinicians. The literature is complex, with a variety of study designs, often of poor quality, used to appraise multiple iterations of different tools. While the pooled accurate diagnosis retrieval rates, particularly for the newer versions, are high; the review suggests that the benefit may be less in complex cases, when they are most likely to be used. Further prospective research is required to demonstrate clinical effectiveness.

What is already known on this topic

Diagnostic errors commonly occur and can lead to preventable patient harm. Their causes are multifactorial and previous efforts to address diagnostic errors have focused on training clinicians to improve clinical and cognitive skills. Electronic DDX generators are evolving technologies which have the potential to reduce error by augmenting and influencing the diagnostic reasoning process of clinicians.

What this study adds

Conclusions are limited by the heterogeneous study designs and poor study quality. DDX generators generally report high levels of accurate diagnosis retrieval but the clinical relevance of this term is questionable and there is no evidence that they perform significantly better than clinicians. However, DDX generators are not intended to replace clinicians and the relatively high levels of accurate diagnosis retrieval observed may augment the decision-making process of generalist clinicians. We also offer insights into what future studies should entail. Firstly, the integrity of the internal validity of studies must remain robust and primary studies must

be conducted independently with no competing interests. Also, in addition to the efficacy of DDX tools, there are a number of important variables including the ranking of diagnoses, cost-effectiveness and time taken to use which should all be addressed in future research studies before further recommendations can be made.

Supporting Information

S1 PRISMA Checklist. PRISMA checklist.

doi:10.1371/journal.pone.0148991.s001
(DOC)

S1 File. QUADAS-2 signalling questions.

doi:10.1371/journal.pone.0148991.s002
(DOCX)

S2 File. Risk of bias of individual studies.

doi:10.1371/journal.pone.0148991.s003
(DOCX)

S3 File. Utility data summary.

doi:10.1371/journal.pone.0148991.s004
(DOCX)

Acknowledgments

We thank Professor Anne-Marie Glenney for her input and advice during this review.

Transparency: NR affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Author Contributions

Conceived and designed the experiments: NR MP RA SCS SC AE PB. Performed the experiments: NR MP RA SCS. Analyzed the data: NR MP RA SCS. Contributed reagents/materials/analysis tools: NR MP RA SCS SC AE PB. Wrote the paper: NR MP RA SCS.

References

1. Balla J, Heneghan C, Goyder C, Thompson M. Identifying early warning signs for diagnostic errors in primary care: a qualitative study. *BMJ Open*. 2012;2:1–9. doi: 10.1136/bmjopen-2012-001539 pmid:22983786.
View Article • PubMed/NCBI • Google Scholar
2. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *The American journal of medicine*. 2008;121(5 Suppl):S2–23. doi: 10.1016/j.amjmed.2008.01.001 pmid:18440350.
View Article • PubMed/NCBI • Google Scholar
3. Kostopoulou O, Oudhoff J, Nath R, Delaney BC, Munro CW, Harries C, et al. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. *Med Decis Making*. 2008;28(5):668–80. doi: 10.1177/0272989X08319958 pmid:18556634.
View Article • PubMed/NCBI • Google Scholar
4. Schiff GD, Puopolo AL, Huben-Kearney A, Yu W, Keohane C, McDonough P, et al. Primary care closed claims experience of Massachusetts malpractice insurers. *JAMA internal medicine*. 2013;173(22):2063–8. Epub 2013/10/02. doi: 10.1001/jamainternmed.2013.11070 pmid:24081145.
View Article • PubMed/NCBI • Google Scholar
5. Graber ML, Kissam S, Payne VL, Meyer AN, Sorensen A, Lenfestey N, et al. Cognitive interventions to reduce diagnostic error: a narrative review. *BMJ*

- Qual Saf. 2012;21(7):535–57. doi: 10.1136/bmjqs-2011-000149 pmid:22543420.
View Article • PubMed/NCBI • Google Scholar
6. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med.* 2003;78(8):775–80. doi: 10.1097/00001888-200308000-00003.
View Article • PubMed/NCBI • Google Scholar
7. Neurath PW, Enslein K, Mitchell GW Jr. Design of a computer system to assist in differential preoperative diagnosis for pelvic surgery. *The New England journal of medicine.* 1969;280(14):745–9. Epub 1969/04/03. doi: 10.1056/NEJM196904032801403 pmid:5773355.
View Article • PubMed/NCBI • Google Scholar
8. Parodi S, Filiberti R, Marroni P, Libener R, Ivaldi G, Mussap M, et al. Differential diagnosis of pleural mesothelioma using Logic Learning Machine. *BMC Bioinformatics.* 2015;16(Suppl 9):S3. doi: 10.1186/1471-2105-16-S9-S3. pmid:26051106
View Article • PubMed/NCBI • Google Scholar
9. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: An evaluation of currently available computer programs. *Journal of General Internal Medicine.* 2011;27(2):213–9. doi: 10.1007/s11606-011-1804-8.
View Article • PubMed/NCBI • Google Scholar
10. Garg A, Adhikari N, Beyene J, Sam J, Haynes R. Effects of computerized clinical decision support systems on practitioner performance. *JAMA.* 2005;293(10):1223–38. pmid:15755945
View Article • PubMed/NCBI • Google Scholar
11. El-Kareh R, Hasan O, Schiff G. Use of health information technology to reduce diagnostic errors. *BMJ Quality and Safety.* 2013;22(2):1–12.
View Article • PubMed/NCBI • Google Scholar
12. Foot C, Naylor C, Imison C. The quality of GP diagnosis and referral. 2010.
13. Graber MA, VanScoy D. How well does decision support software perform in the emergency department? *Emergency Medicine Journal.* 2003;20(5):426–8.
View Article • PubMed/NCBI • Google Scholar
14. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology.* 2009;62(10):e1–34. Epub 2009/07/28. doi: 10.1016/j.jclinepi.2009.06.006 pmid:19631507.
View Article • PubMed/NCBI • Google Scholar
15. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. *New England Journal of Medicine.* 1994;330(25):1792–6. doi: 10.1056/NEJM199406233302506.
View Article • PubMed/NCBI • Google Scholar
16. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
View Article • PubMed/NCBI • Google Scholar
17. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529–36. doi: 10.7326/0003-4819-155-8-201110180-00009. pmid:22007046
View Article • PubMed/NCBI • Google Scholar
18. Borenstein M. *Comprehensive meta-analysis.* Englewood, NJ: Biostat. 2005.
19. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. *Jama.* 1999;282(19):1851–6. Epub 1999/11/26. pmid:10573277.
View Article • PubMed/NCBI • Google Scholar
20. Heckerling PS, Elstein AS, Terzian CG, Kushner MS. The effect of incomplete knowledge on the diagnoses of a computer consultant system. *Medical Informatics.* 1991;16(4):363–70.
View Article • PubMed/NCBI • Google Scholar

21. Murphy GC, Friedman CP, Elstein AS, Wolf FM, Miller T, Miller JG. The influence of a decision support system on the differential diagnosis of medical practitioners at three levels of training. *Proceedings: a conference of the American Medical Informatics Association AMIA Annual Fall Symposium AMIA Fall Symposium*. 1996;219–23. pmid:8947660.
View Article • PubMed/NCBI • Google Scholar
22. Ramnarayan P, Roberts GC, Coren M, Nanduri V, Tomlinson A, Taylor PM, 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;6:22. Epub 2006/05/02. doi: 10.1186/1472-6947-6-22 pmid:16646956; PubMed Central PMCID: PMC1513379.
View Article • PubMed/NCBI • Google Scholar
23. 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 Medical Informatics and Decision Making*. 2006;6(1):37. doi: 10.1186/1472-6947-6-37.
View Article • PubMed/NCBI • Google Scholar
24. Sedgwick P, Marston L. Meta-analyses: standardised mean differences 2013 2013-12-06 11:36:16.
25. Durlak J. How to Select, Calculate, and Interpret Effect Sizes. *Journal of Pediatric Psychology*. 2009;34(9):917–28. doi: 10.1093/jpepsy/jsp004. pmid:19223279
View Article • PubMed/NCBI • Google Scholar
26. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*. 2003;327(7414):557–60. PMC192859. pmid:12958120
View Article • PubMed/NCBI • Google Scholar
27. Egger M, Davey-Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629–34. pmid:9310563
View Article • PubMed/NCBI • Google Scholar
28. Pan N, Leoncini E, de Belvis G, Ricciardi W, Boccia S. Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. *PloS one*. 2013;8(12):e83138. Epub 2014/01/05. doi: 10.1371/journal.pone.0083138 pmid:24386151; PubMed Central PMCID: PMC3873291.
View Article • PubMed/NCBI • Google Scholar
29. Apkon M, Mattera JA, Lin Z, Herrin J, Bradley EH, Carbone M, et al. A randomized outpatient trial of a decision-support information technology tool. *Archives of internal medicine*. 2005;165(20):2388–94. CN-00531948 UPDATE. pmid:16287768
View Article • PubMed/NCBI • Google Scholar
30. Arene I, Ahmed W, Fox M, Barr CE, Fisher K. Evaluation of quick medical reference (QMR) as a teaching tool. *MD Comput: computers in medical practice*. 1998;15(5):323–6. pmid:9753979.
View Article • PubMed/NCBI • Google Scholar
31. Bacchus CM, Quinton C, O'Rourke K, Detsky AS. A randomized crossover trial of quick medical reference (QMR) as a teaching tool for medical interns. *Journal of General Internal Medicine*. 1994;9(11):616–21.
View Article • PubMed/NCBI • Google Scholar
32. Bankowitz RA, McNeil MA, Challinor SM, Parker RC, Kapoor WN, Miller RA. A computer-assisted medical diagnostic consultation service. Implementation and prospective evaluation of a prototype. *Annals of Internal Medicine*. 1989;110(10):824–32.
View Article • PubMed/NCBI • Google Scholar
33. Berner ES, Maisiak RS, Cobbs CG, Taunton OD. Effects of a decision support system on physicians' diagnostic performance. *Journal of the American Medical Informatics Association*. 1999;6(5):420–7.
View Article • PubMed/NCBI • Google Scholar
34. Carlson J, Abel M, Bridges D, Tomkowiak J. The impact of a diagnostic reminder system on student clinical reasoning during simulated case studies. *Simulation in healthcare: Journal of the Society for Simulation in Healthcare*. 2011;6(1):11–7. pmid:21330845.
View Article • PubMed/NCBI • Google Scholar

35. Elkin PL, Liebow M, Bauer BA, Chaliki S, Wahner-Roedler D, Bundrick J, et al. The introduction of a diagnostic decision support system (DXplain) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging Diagnostic Related Groups (DRGs). *International Journal of Medical Informatics*. 2010;79(11):772–7. doi: 10.1016/j.ijmedinf.2010.09.004.
View Article • PubMed/NCBI • Google Scholar
36. Elstein AS, Friedman CP, Wolf FM, Murphy G, Miller J, Fine P, et al. Effects of a decision support system on the diagnostic accuracy of users: a preliminary report. *Journal of the American Medical Informatics Association: JAMIA*. 1996;3(6):422–8. pmid:8930858.
View Article • PubMed/NCBI • Google Scholar
37. Graber ML. Taking steps towards a safer future: measures to promote timely and accurate medical diagnosis. *The American journal of medicine*. 2008;121(5 Suppl):S43–6. doi: 10.1016/j.amjmed.2008.02.006 pmid:18440355.
View Article • PubMed/NCBI • Google Scholar
38. Graber ML, Tompkins D, Holland JJ. Resources medical students use to derive a differential diagnosis. *Med Teach*. 2009;31(6):522–7. Epub 2009/10/09. pmid:19811168.
View Article • PubMed/NCBI • Google Scholar
39. Hammersley JR, Cooney K. Evaluating the Utility of Available Differential Diagnosis Systems. *Proceedings of the Annual Symposium on Computer Application in Medical Care*. 1988:229–31.
View Article • PubMed/NCBI • Google Scholar
40. Lange LL, Haak SW, Lincoln MJ, Thompson CB, Turner CW, Weir C, et al. Use of Iliad to improve diagnostic performance of nurse practitioner students. *The Journal of nursing education*. 1997;36(1):36–45. pmid:8986960.
View Article • PubMed/NCBI • Google Scholar
41. Lemaire JB, Schaefer JP, Martin LA, Faris P, Ainslie MD, Hull RD. Effectiveness of the quick medical reference as a diagnostic tool. *Cmaj*. 1999;161(6):725–8.
View Article • PubMed/NCBI • Google Scholar
42. Li YC, Haug PJ, Lincoln MJ, Turner CW, Pryor TA, Warner HH. Assessing the behavioral impact of a diagnostic decision support system. *Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care*. 1995:805–9. pmid:8563402.
View Article • PubMed/NCBI • Google Scholar
43. Miller RA, McNeil MA, Challinor SM. The INTERNIST-1/QUICK MEDICAL REFERENCE Project—Status report. *Western Journal of Medicine*. 1986;145(6):816–22.
View Article • PubMed/NCBI • Google Scholar
44. Nelson SJ, Blois MS, Tuttle MS, Erlbaum M, Harrison P, Kim H, et al. Evaluating RECONSIDER. A computer program for diagnostic prompting. *J Med Syst*. 1985;9(5–6):379–88. Epub 1985/12/01. pmid:3912452.
View Article • PubMed/NCBI • Google Scholar
45. Ramnarayan P, Cronje N, Brown R, Negus R, Coode B, Moss P, et al. Validation of a diagnostic reminder system in emergency medicine: A multi-centre study. *Emergency Medicine Journal*. 2007;24(9):619–24. doi: 10.1136/emj.2006.044107.
View Article • PubMed/NCBI • Google Scholar
46. Waxman HS, Worley WE. Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer-based system. *Medicine*. 1990;69(3):125–36. pmid:2189054.
View Article • PubMed/NCBI • Google Scholar
47. Wexler JR, Swender PT, Tunnessen WW Jr, Oski FA. Impact of a system of computer-assisted diagnosis. Initial evaluation of the hospitalized patient. *American Journal of Diseases of Children*. 1975;129(2):203–5. CN-00217040. pmid:1091140
View Article • PubMed/NCBI • Google Scholar
48. Wolf FM, Friedman CP, Elstein AS, Miller JG, Murphy GC, Heckerling P, et al. Changes in diagnostic decision-making after a computerized decision support consultation based on perceptions of need and helpfulness: a preliminary report. *Proceedings: a conference of the American Medical Informatics Association AMIA Annual Fall Symposium AMIA Fall Symposium*. 1997:263–7. pmid:9357629.
View Article • PubMed/NCBI • Google Scholar

49. Bavdekar SB, Pawar M. Evaluation of an Internet delivered pediatric diagnosis support system (ISABEL) in a tertiary care center in India. *Indian pediatrics*. 2005;42(11):1086–91. pmid:16340049.
View Article • PubMed/NCBI • Google Scholar
50. Feldman MJ, Barnett GO. An approach to evaluating the accuracy of DXplain. *Computer Methods & Programs in Biomedicine*. 1991;35(4):261–6. pmid:1752121.
View Article • PubMed/NCBI • Google Scholar
51. Gozum ME. Emulating cognitive diagnostic skills without clinical experience: a report of medical students using Quick Medical Reference and Iliad in the diagnosis of difficult clinical cases. *Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care*. 1994;991. pmid:7950096.
View Article • PubMed/NCBI • Google Scholar
52. Lau LM, Warner HR, Poulsen AL. Research review: a computer-based diagnostic model for individual case review. *Topics in health information management*. 1995;15(3):67–79. pmid:10140306.
View Article • PubMed/NCBI • Google Scholar
53. Lincoln MJ, Turner CW, Haug PJ, Warner HR, Williamson JW, Bouhaddou O, et al. Iliad training enhances medical students' diagnostic skills. *Journal of Medical Systems*. 1991;15(1):93–110. doi: 10.1007/BF00993883.
View Article • PubMed/NCBI • Google Scholar
54. Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *The New England journal of medicine*. 1982;307(8):468–76. Epub 1982/08/19. doi: 10.1056/nejm198208193070803 pmid:7048091.
View Article • PubMed/NCBI • Google Scholar
55. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: A web-based differential diagnostic aid for paediatrics: Results from an initial performance evaluation. *Archives of Disease in Childhood*. 2003;88(5):408–13. doi: 10.1136/adc.88.5.408.
View Article • PubMed/NCBI • Google Scholar
56. Rodriguez-Gonzalez A, Torres-Nino J, Mayer MA, Alor-Hernandez G, Wilkinson MD. Analysis of a multilevel diagnosis decision support system and its implications: A case study. *Computational and Mathematical Methods in Medicine*. 2012;2012(367345). doi: 10.1155/2012/367345.
View Article • PubMed/NCBI • Google Scholar
57. 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. doi: 10.1007/s11606-007-0271-8 pmid:18095042; PubMed Central PMCID: PMC2150633.
View Article • PubMed/NCBI • Google Scholar
58. Bankowitz R, Blumenfeld B, Guise Bettinsoli N. User variability in abstracting and entering printed case histories with Quick Medical Reference (QMR). *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care*. 1987:68–73.
View Article • PubMed/NCBI • Google Scholar
59. Bond W, Schwartz L, Weaver K, Levick D, Giuliano M, Graber M. A qualitative review of differential diagnosis generators. Poster presented at: The 32nd Annual Meeting of the Society for Medical Decision Making, Ontario, Canada. 2010.
60. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ*. 2000;321(7262):694–6. doi: 10.1136/bmj.321.7262.694. pmid:10987780
View Article • PubMed/NCBI • Google Scholar
61. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet* 2000;356(9237):1228–31. pmid:11072941
View Article • PubMed/NCBI • Google Scholar
62. Kontopantelis E, Springate DA, Reeves D. A Re-Analysis of the Cochrane Library Data: The Dangers of Unobserved Heterogeneity in Meta-Analyses. *PLoS one*. 2013;8(7):e69930. doi: 10.1371/journal.pone.0069930. pmid:23922860
View Article • PubMed/NCBI • Google Scholar