Automation bias: a systematic review of frequency, effect mediators, and mitigators

Kate Goddard,[1] Abdul Roudsari,[1,2] Jeremy C Wyatt[3]

► Additional materials are published online only. To view these files please visit the journal online (www.jamia.org/content/19/1.toc).

[1]Centre for Health Informatics, City University, London, UK
[2]School of Health Information Science, University of Victoria, Victoria, British Columbia, Canada
[3]Institute of Digital Healthcare, University of Warwick, UK

Correspondence to
Kate Goddard, Centre for Health Informatics, City University, Northampton Square, London EC1V 0HB, UK;
kate.goddard.1@city.ac.uk

Received 17 December 2010
Accepted 17 May 2011
Published Online First
16 June 2011

ABSTRACT

Automation bias (AB)—the tendency to over-rely on automation—has been studied in various academic fields. Clinical decision support systems (CDSS) aim to benefit the clinical decision-making process. Although most research shows overall improved performance with use, there is often a failure to recognize the new errors that CDSS can introduce. With a focus on healthcare, a systematic review of the literature from a variety of research fields has been carried out, assessing the frequency and severity of AB, the effect mediators, and interventions potentially mitigating this effect. This is discussed alongside automation-induced complacency, or insufficient monitoring of automation output. A mix of subject specific and freetext terms around the themes of automation, human—automation interaction, and task performance and error were used to search article databases. Of 13 821 retrieved papers, 74 met the inclusion criteria. User factors such as cognitive style, decision support systems (DSS), and task specific experience mediated AB, as did attitudinal driving factors such as trust and confidence. Environmental mediators included workload, task complexity, and time constraint, which pressurized cognitive resources. Mitigators of AB included implementation factors such as training and emphasizing user accountability, and DSS design factors such as the position of advice on the screen, updated confidence levels attached to DSS output, and the provision of information versus recommendation. By uncovering the mechanisms by which AB operates, this review aims to help optimize the clinical decision-making process for CDSS developers and healthcare practitioners.

BACKGROUND

Clinical decision support systems (CDSS) have great potential to improve clinical decisions, actions, and patient outcomes[1 2] by providing advice and filtered or enhanced information, or by presenting prompts or alerts to the user. However, most studies of CDSS have emphasized the accuracy of the computer system, without placing clinicians in the role of direct users. Although most decision support systems (DSS) are 80–90% accurate, it is known that the occasional incorrect advice they give may tempt users to reverse a correct decision they have already made, and thus introduce errors of over-reliance.[3] These errors can be classified as *automation bias* (AB),[4] by which users tend to over-accept computer output 'as a heuristic replacement of vigilant information seeking and processing.'[5] AB manifests in errors of commission (following incorrect advice) and omission (failing to act because of not being prompted to do so) when using CDSS. Indeed, the implementation of interventions such as CDSS can actually introduce new errors.[6] However, AB has not previously been properly defined or examined; the negative effects need investigation particularly in the healthcare field, where decision errors can have severe consequences. It is currently unclear how frequently AB occurs and what the risk factors are and so a systematic review of the current literature was carried out to clarify the nature of AB.

Automation bias and complacency

Previous investigation of AB has primarily focused on the aviation sector. The examination of human factors involved in healthcare systems is more recent, and no systematic reviews of the reliance (over-reliance in particular) phenomenon relating to healthcare have been identified so far. In a study examining the enhancement of clinical decision-making by the use of a computerized diagnostic system, Friedman *et al*[7] noticed that in 6% of cases, clinicians over-rode their own correct decisions in favor of erroneous advice from the DSS. Despite the high risk associated with medication error, there is little direct research into over-reliance on technology in the healthcare and CDSS fields.

Often the literature has looked solely at overall clinical or DSS accuracy and clinical outcomes without investigating etiology and types of errors. More recent papers have started to examine the human factors relevant to appropriate design and use of automation in general—for example, trust[8] and other social, cognitive, and motivational factors. As the concept is relatively new and undefined, a number of synonyms have been used in the literature to describe the concept of AB, including automation-induced complacency,[9] over-reliance on automation, and confirmation bias.[10]

In a recent literature review, Parasuraman *et al*[11] discuss AB and automation-induced complacency as overlapping concepts reflecting the same kind of automation misuse associated with misplaced attention: either an attentional bias toward DSS output, or insufficient attention and monitoring of automation output (particularly with automation deemed reliable). Parasuraman *et al* note that commission and omission errors result from both AB and complacency (although they mention that commission errors are more strictly the domain of AB). There is a lack of consensus over the definition of complacency. Complacency appears to occur as an attention allocation strategy in multitasking where manual tasks are attended to over monitoring the veracity of the output of automation. Automation bias can also be found outside of

multitask situations, and occurs when there is an active bias toward DSS in decision-making.

Although the focus of this review is on AB, due to the theoretical overlap and vagaries with current definitions, it will also include papers which imply complacency effects as these also affect the misuse of or over-reliance on automation literature. Similar outcomes in terms of commission or omission may mean that one effect may be conflated by or confused with another. Studies relating to AB are identified and examined separately from those on automation complacency.

REVIEW AIM AND OBJECTIVES

The overall aim is to systematically review the literature on DSS and AB, particularly in the field of healthcare. The specific review objectives are to answer the following questions:

- What is the rate of AB and what is the size of the problem?
- What are the mediators for AB?
- Is there a way to mitigate AB?

PRISMA methodology[12] was used to select articles, and involved identification, screening, appraisal for eligibility, and qualitative and quantitative assessment of final papers.

REVIEW METHODS

Sources of studies

The main concepts surrounding the subject of AB are DSS intervention, the DSS—human interaction, and task performance and error generation. These were the key themes used in a systematic search of the literature. As initial searches indicated little healthcare specific evidence, it was decided to include a number of databases and maintain wide parameters for inclusion/exclusion to identify further relevant articles.

The search took place between September 2009 and January 2010 and the following databases were searched: MEDLINE/PubMed, CINAHL, PsycInfo, IEEE Explore, and Web of Science.

No timeframe limit was set for any database and the language filter was set to English language studies only.

According to the three concepts illustrated in figure 1, MeSH search terms, context specific freetext search terms, and non-MeSH index terms were combined. The details are available on request. The field of healthcare was the focus for the search, but papers from any discipline were considered.

Eligibility criteria for studies

It was clear from preliminary searches that the review should not be limited to a specific field. Investigation of decision support and automation in non-healthcare disciplines can supply valuable information on human—computer systems and cognitive biases, and provide recommendations on how to debias individuals. The exploratory nature of the research justified a broad, multi-disciplinary search in order to identify the widest cross-section of papers. All study settings and all papers irrespective of the level of the user's expertise were considered.

The first search of databases identified 14 457 research papers (inclusive of duplicates).

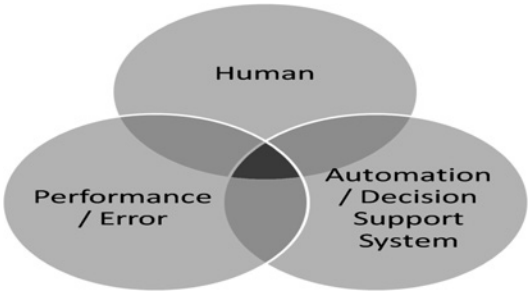

Figure 1 Diagram of search concepts.

Inclusion criteria

Papers were included which investigated:

- human interaction with automated DSS in any field (eg, aviation, transport, and cognitive psychology), but particularly healthcare.
- empirical data on automation use, especially those which incorporated a subjective user questionnaire or interview.
- the appropriateness and accuracy of use of DSS.

Particular attention was given to papers explicitly mentioning AB, automation misuse, or over-reliance on automation, or terms such as confirmation bias, automation dependence, or automation complacency.

Outcome measures

Papers were included which reported:

- assessment of user performance (the degree and/or appropriateness of automation use), which included:
- indicators of DSS advice usage—consistency between DSS advice and decisions, user performance, peripheral behaviors (such as advice verification), or response bias indicators.[13]
- indicators of the influence of automation on decision-making, such as pre- and post-advice decision accuracy (eg, negative consultations, where a pre-advice decision is correct and changed to an incorrect post-advice decision), DSS versus non-DSS decision accuracy (higher risk of incorrect decisions when bad advice is given by DSS vs control group decisions), and correlation between DSS and user accuracy (the relationship between falling DSS accuracy and falling user decision accuracy), such as user sensitivity[i] and specificity[ii] with varying DSS accuracy.
- analysis of error types (such as those of commission or omission) and reasons for user error, or ineffective DSS use.

Screening and eligibility

See online supplementary appendix 1 for the stages of the literature retrieval process. Obvious duplicate articles were removed at stage 1 and the remainder were removed at stage 2 using Endnote and a visual search.

Reliability

When articles were extracted from a pseudo-randomized sample of papers, the crude rate of agreement between two reviewers was 87%. Cohen's kappa was 0.8436; according to Landis and Koch[14] who formulated a table for assessing kappa significance, this result implies 'almost perfect agreement.'

Quality assessment

Once the full article had been read, the papers were scored for internal and external validity and for relevance to the review aims. Generic paper quality was scored according to items adapted from the PRISMA CONSORT checklist. Paper relevance was scored according to the outcome measures and inclusion criteria and structured according to Population, Intervention,

[i]Sensitivity measures the proportion of correctly identified true positives; higher sensitivity is related to a lower false negative rate.
[ii]Specificity measures the proportion of negatives which are correctly identified; higher specificity is related to a lower false positive rate.

Control, and Outcome (PICO). Three papers were removed from the final sample because they were not relevant to the final study. Quality assessment was kept fairly flexible—Juni *et al*[15] advise against overly rigorous scoring and selection of studies based on a threshold, advocating that papers should be selected on their individual merits to avoid skewing results.

Data extraction

The results were organized and tabulated according to an extended PICO framework (see online supplementary table) according to:

- Population: all users, any demography, background, or level of expertise
- Design
- Intervention/exposure
- Comparison: groups not using automated decision support, or different forms of decision support (non-automated or automated but a different design), or before and after design
- Outcome measures: assessment of user performance; error types and reasons for user error
- Other relevant information.

FINDINGS

From an initial 13 821 papers (after removal of duplicates), a total of 74 studies were found (table 1) which satisfied the inclusion/exclusion criteria. The concept of AB was first discussed and continues to be explored most fully in the aviation field.

The main areas of clinical study are based around computer aided detection (CAD) type DSS[16] followed by ECG use.[17] Two studies investigating general diagnostic-based CDSS (eg, QMR, Iliad) were found.[7 18] Other DSS include more domain-specific DSS related to ECG reading,[17 19 20] skin lesions,[21] antibody identification,[22] and chest pain.[23]

Table 1 Profile of papers found in a systematic review by research field and year of publication

	Healthcare			Generic			
	CAD	Other	Aviation	HCI	Military	Other	Total
1993–1996	1	3	6	1	0	0	11
1997–2000	1	4	8	0	0	1	14
2001–2004	7	3	6	4	1	2	23
2005–2008	6	5	7	1	2	4	25
2009		1					1
Total	15	16	27	6	3	7	74

CAD, computer aided detection; HCI, Human-Computer Interaction.

DISCUSSION

The rate and extent of AB

Automation bias appears to be a fairly robust and generic effect across research fields. Most studies found that DSS improved user performance overall, even when the advice given was inappropriate,[22] although some showed overall decreases in performance.[23]

Regarding outcome measures, errors relating to erroneous DSS output were recognized in terms of negative consultations,[7 18 20 24 25] percentage of erroneous advice cases followed,[17 19 26 27] and more indirect implied measures of AB such as a decrease in accuracy when DSS is inaccurate,[23 28–30] or if there is a correlation between decreasing DSS accuracy and decreasing user accuracy.[31–33]

Studies on CAD focused on AB effects, showing mixed results and distinguishing between errors of commission and omission as regards sensitivity and specificity. Four studies showed a decrease in both measures with inaccurate DSS due to AB.[16 34–36] Four reports showed contrasting effects on sensitivity and specificity, finding increased sensitivity with a decrease in specificity with CAD.[37–40] It was found that CAD interventions can decrease reported specificity without decreasing sensitivity[35 39 41] but have also been found to increase specificity with no effect on sensitivity.[42] Some studies explicitly state that no AB was found despite there being the opportunity for it to emerge.[42–44] Moberg *et al*[42] stated this was mostly due to false-positive targets detected with CAD output being generally different from those detected by human observers, thus it was relatively easy for observers to disregard false-positives, the potential cost of higher automation error being mitigated by pilot strategy, whereby the sounding of an alert led to a closer scrutiny of the raw data.

Four papers in the healthcare field found that user accuracy decreased with erroneous DSS intervention in comparison with a non-intervention control group. Results from these four papers were pooled in a small, indicative meta-analysis on the basis that they assessed the percentage of erroneous decisions following incorrect advice given by CDSS compared to a non-CDSS control. The studies were homogeneous in terms of methodology, control group, intervention type, and field of study and had high quality scores. The CDSS analyzed were non-interruptive in nature and the advice text-based. The studies also analyzed commission errors, which are more clearly AB than complacency errors. These were included in a Mantel–Haenszel (M–H) method random effects model, risk ratio analysis,[iii] at the 95% confidence level. Studies are summarized in the online supplementary appendix 2. The risk ratio was 1.26 (95% CI 1.11 to 1.44); erroneous advice was more likely to be followed in the CDSS groups than in the control groups and when in error the CDSS increased the risk of an incorrect decision being made by 26%.[iv] The RevMan program was used to analyze the papers and the results are shown in figure 2.

Non-controlled effects—negative consultations

Negative consultations are the clearest measure of AB, for example, as compared to the percentage of incorrect decisions following incorrect advice, which could be conflated by users having the same incorrect pre-advice decision (thus no AB-generated decision 'switching' will have occurred, despite being included in the calculation). Four studies with similar designs reported on this outcome. The proportion of decisions which demonstrated this ranged from 6%[7] to 11%[25] of cases in prospective empirical studies.

Friedman *et al*[7] found positive consultations, where the correct diagnosis was present after consultation but not before, in 12% of cases; negative consultations were observed in 6% of cases. The resultant net gain was 6%. Berner *et al*[18] found that in 21 cases (of 272) the correct unaided answer was changed to an incorrect answer after DSS use; 8% were negative consultations. Westbrook *et al*[24] found that system use resulted in a 21% improvement in clinicians' answers, from 29% correct before system use to 50% after system use, however 7% of correct

[iii]The *fixed effects model* assumes that all studies are based on a common population, and that the effect size (OR) is not significantly different among the different trials. The *random effects assumption* (made in a random effects model) is that the individual specific effects are uncorrelated with the independent variables.
[iv]The overall effect was significant, $p<0.0005$. Tests for heterogeneity were not significant ($p>0.05$), implying that the variation in underlying intervention effects across studies was not significant.

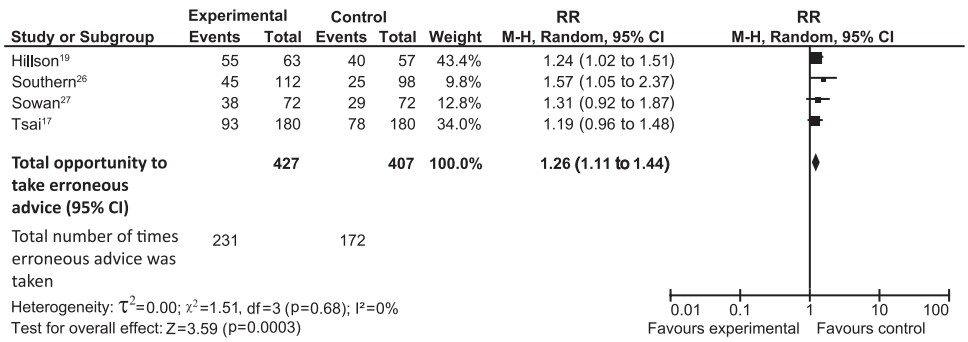

Figure 2 RevMan meta-analysis output of four papers showing erroneous advice followed (of total opportunities).

pre-use answers were changed incorrectly. A similar study by McKibbon and Fridsma,[25] which examined how clinician-selected electronic information resources improved physicians' answers to simulated clinical questions, found a negative consultation rate of 11%.

Interruptive DSS in aviation

Interruptive DSS studies into AB were mainly found in the field of aviation research. Skitka *et al* found commission errors to be higher than omission errors in two studies into AB,[4 45] while Mosier *et al*[46] found 55% omission rates and 0% commission rates in an aviation study. Many reports did not distinguish between omission and commission rates, reporting overall errors only. Automation complacency error rates for interruptive systems have been shown to increase if a DSS is highly (but not perfectly) reliable, leading to overtrust and complacency[47–50] and to decrease if it is less reliable (but not highly, obviously unreliable). Lower levels of reliability can paradoxically inspire better performance due to lower complacency levels,[49] for example, Madhavan and Wiegmann[51] set the optimal threshold at 70% reliability before performance degrades or the DSS is disused.

Causes of AB

When AB is reported, it ranges from a significant problem[17 19] which may render DSS unhelpful on balance (according to the rate and severity of automation error) to a lesser problem where it is still worthwhile given the benefits.[40] Investigation of potential effect modifiers is crucial for understanding the underlying causes of AB.

Experience

General and DSS-specific *experience* has been shown to affect the tendency toward over-reliance in eight papers. Reports which focus on complacency[52] and AB[53] suggest that task inexperience may lead to automation-related errors, however inexperienced users showed the best overall improvement from using DSS.[39] If user have been trained, experience may decrease over-reliance on automation by different mechanisms; in complacency it may familiarize users with baseline reliability, and in AB it may highlight the risk of accepting incorrect information, promoting verification of uncertain output. Walsham *et al*[54] showed that despite no apparent improvement in performance, CAD improved the subjective confidence of one less experienced user, while it improved only the overall performance of less versus more experienced users; CAD can lead to a mismatch of decision confidence against actual performance, and is of greater value to users with less task experience. Automation bias occurs more often with task inexperienced users[53 55] but can occur with more experienced users.[39] Physicians with greater experience may be less reliant on DSS and be more likely to recognize incorrect advice.[7 21 53] In an experiment examining reliance on medication management systems, Ho *et al*[30] found that age was a factor in DSS related error, with older users making more errors, although this may be an indirect relationship mediated by experience. Conversely, regarding complacency, Bailey and Scerbo[47] found that specific DSS experience decreased monitoring of performance—familiarity led to desensitization and habituation effects.

Confidence and trust

Experience may be positively related to *user confidence*. Three papers reporting the results of multitask experiments, showed that increased confidence[47 49 56] in the user's own decision decreased reliance on external support, whereas trust in the DSS increased reliance.[56] Similarly, Dreiseitl and Binder[21] showed that physicians were more likely to be biased by automation and accept DSS advice when they were less confident of their own diagnosis. Lee and Moray[57] state that automation reliance is essentially a trade-off between self-confidence and trust in the DSS. The relationship between *trust*[30 58] and automation reliance has arguably been the subject of most research on complacent behaviors[47–49 59 60] and AB.[51 56 61–63] Trust is possibly the strongest driving factor in over-reliance, when trust is incorrectly calibrated against system reliability. This may be a general trend in human judgment, for example, Dzindolet *et al*[62] demonstrated that users had a predisposition to trust, or had a 'positivity bias' toward, a automated aid over a human one and commit AB error. Higher perceived automation pedigree[51] (for example novice vs expert systems) also affects reliance, increasing trust in the system.

Individual differences

Individual differences in reliance have been found for example, to affect the potential for complacency[9 47 59] and also for predilection for certain decision strategies such as maximization[63] or non-compensatory decision strategies (vs compensatory) which use minimal information on which to base decisions.[64] Underlying personality and cognitive characteristics may predispose some users to committing automation based errors in terms of both AB[30 46 61 63 65] and complacency.[9 59] Producing DSS which provide good cognitive fit could decrease AB error rates.

Task type

The *task type* itself may affect how users rely on external automated advice. More complex tasks and higher workloads are posited to increase reliance[47] by placing stress on cognitive capacity. Users may become biased to overuse automated advice under increased workload[61] or may be prone to automation complacency.[66–68] Xu *et al*,[44] however, found that in contrast to this, increased trial difficulty improved performance, suggesting it decreased user complacency and led to closer inspection of the data. Factors which increase external pressures on an individual's

cognitive capacity may shift reliance toward external support. Prinzel *et al*[59] found that the relationship between perceived workload and complacency error was mediated by users' intrinsic complacency potential. Those with a high complacency potential were more likely to report higher perceived workload and have lower monitoring performance. Sarter and Schroeder[55] suggested that high *time pressure* could bias a user toward DSS usage. Both AB and complacency errors are thought to arise from reallocation of attention[11]; putting pressure on cognitive resources could either bias a user toward heuristically using DSS output, or over-relying on automation to provide correct output so attention can be channeled toward other tasks.

These factors both place stress on cognitive capacity. As an adaptive measure users tend to then rely on DSS to compensate; if the DSS is reliable this is performance enhancing, if not, it can lead to new errors.

Automation bias avoidance

Implementation factors

Research indicates that certain measures can help prevent over-reliance on decision support. One study found that making users aware of the *DSS reasoning process*[62] increased appropriate reliance, thus reducing AB. Increasing *accountability* for decisions may also prevent AB. However, while two studies[45 65] showed that external manipulation of accountability increased vigilance and thus decreased AB, another study[46] showed that external manipulations did not have this affect but that users' internal perceptions of accountability did—people who perceived themselves to be accountable made less AB errors. Similarly, one study[32] found a positive relationship between DSS misuse and negative attitudes in a workplace and shared social norms; therefore, improving the *working culture* may help appropriate DSS use. Papers have assessed the effect of *training* on appropriate DSS use; linked to experience discussed above, training may increase the likelihood of recognizing DSS error and thus reduce AB (particularly commission errors).[45] However, Parasuraman *et al* cite a Mosier *et al*[69] study which implied training had no impact on AB. Complacency error[52 68 70] is more clearly reduced by training than AB.

Design of DSS

The *design* of the DSS can affect how users regard advice. To reduce complacency error, adaptive task allocation[71]—varying reliability rather than keeping it constant[9 72]—was found to increase vigilant behavior and improve appropriate reliance. The position of advice on the screen can affect the likelihood of AB. Berner *et al*[18] found that display prominence increased AB, affecting the likelihood of a decision being changed after advice—prominent incorrect advice is more likely to be followed. However, Singh *et al*[73] found that while DSS intervention produced more complacent performance compared to a manual control, centrally (vs peripherally) locating the monitoring task on the screen made no difference to this performance. In another study into automation complacency, Yeh and Wickens[60] investigated system factors; too much on-screen detail makes people less conservative, thus increasing biases. This study also found, conversely, that increasing scene realism appeared to increase conservative decisions. McGuirl and Sarter[56] found that updating the confidence level of the DSS alongside pieces of advice (as opposed to providing one overall fixed confidence level for the system) improved the appropriateness of user reliance, decreasing AB. Sarter and Schroeder[55] suggested that status displays (vs command type displays) rendered imperfect DSS less likely to cause AB—while display helps with detecting a problem, command type advice cuts out a step in the decision-making process and thus may be prone to overuse under time pressure.

Thus there is evidence that AB can be mitigated by decreasing the prominence of DSS output, but there is no evidence for this in complacency, while complacency can be reduced by adaptive task allocation. Automation bias can be reduced by decreasing on-screen detail, updating advice confidence levels, and providing supportive information rather than commands.

OVERALL CONCLUSION

Although research does exist which demonstrates the AB effect, there appear to be few definitive and deliberate studies examining how inaccurate DSS advice affects the user's decision.

There are a number of factors (in terms of user, DSS, task, and environmental characteristics) which may directly or indirectly impact a user's tendency to accept inaccurate advice, and the ways this can be mitigated. The primary drivers for AB and complacency may be user calibration of the trade-off between trust and confidence. This is tempered by individual predisposition in terms of cognitive style and technology acceptance. Task specific and previous DSS experience may act on primary drivers to influence reliance on DSS. Environmental factors such as task complexity and workload, and time pressure can also place pressure on cognitive resources leading to more heuristic-based use of DSS output; if output is incorrect, this can lead to over-reliance. Methods to mitigate AB include implementation and DSS design factors. Increasing user accountability for decisions and DSS training improve appropriate reliance. The presentation of additional information such as up-to-date DSS confidence levels can improve appropriate reliance, as can design factors such as the position of advice on the screen and mode of advice (for example, information vs recommendation). Parasuraman and Manzey[11] carried out a broad review of the literature, including theoretical and anecdotal papers, outlining complacency and AB in several research fields. The focus and scope of this review systematically expands on empirical evidence for AB rates, causes, and mitigators within the healthcare field.

Many factors and complacency effects, which are likely to be interlinked, are involved in AB. Even though the nature of AB is not clear, there are enough studies, discussion papers, and anecdotal evidence to imply that it is a consistent effect. It is postulated frequently but lacks clear empirical evidence.

Limitations

The major unresolved issue is the incidental nature of the reporting of AB. Key papers do not examine this phenomenon and thus it is not mentioned explicitly in their titles, abstracts, or even full text. In addition, both AB and complacency processes remain ill defined; the posited overlap and similarity in error types call for more research on the differences in and relationships between the concepts (such as the integrated model proposed by Parasuraman *et al*[11]). The effect is usually found in a post-hoc analysis of data, and the data reported often show indirect, implicit, evidence of AB. This also means that papers with this finding are likely to have high heterogeneity in their search engine indexing.

Another issue is the heterogeneity of results, which only allowed a smaller meta-analysis to be carried out. Heterogeneity within papers regarding materials, methodology, and outcome

measures, can render direct comparisons difficult.[74] In this instance, systematic review may be best undertaken within the context of a literature review of hypothetical factors to give broader context and meaning to these results.

To address the gaps in empirical evidence due to the anecdotal nature of the available evidence for AB, this review focused on quantitative evidence. However, randomized controlled trials may not be the best method to assess over-reliance on technology in real world settings. Studies based on fieldwork, such as that reported by Campbell *et al*,[75] should be examined in conjunction with more controlled evidence to fully understand the nature of AB.

This review aims to provide an evidence base for the existence of AB. DSS designers, policy makers, implementers, and users should be aware of the nature of automation-induced errors. Given the potentially serious outcomes resulting from incorrect medical decisions, it would be beneficial to examine the negative impact of introducing automated clinical advice, as well as the overall positive effects of CDSS on medical decision-making.

Acknowledgments We would like to thank the *JAMIA* reviewers for their suggestions for improvements and adjustments to this paper.

Funding This review was carried out as part of ongoing PhD research funded by the Engineering and Physical Sciences Research Council.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed

REFERENCES

1. **Garg AX,** Adhikari NK, McDonald H, *et al*. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 2005;**293**:1223—38.
2. **Sullivan F,** Wyatt JC. ABC of Health Informatics 4: how decision support tools help define clinical problems. *BMJ* 2005;**331**:831—3.
3. **Coiera E,** Westbrook J, Wyatt JC. *The Safety and Quality of Decision Support Systems. Yearbook of Medical Informatics 2006*. Stuttgart: Schattauer Verlag, 2006.
4. **Skitka LJ.** Does automation bias decision-making? *Int J Hum Comput Stud* 1999;**51**:991—1006.
5. **Mosier K,** Skitka LJ. Human decision makers and automated decision aids: made for each other? In: Parasuraman R, Mouloua M, eds. *Automation and Human Performance: Theory and Applications*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1996:pp. 201—20.
6. **Koppel R,** Metlay JP, Cohen A, *et al*. Role of computerized physician order entry systems in facilitating medication errors. *JAMA* 2005;**293**:1197—203.
7. **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.
8. **Lee JD,** See KA. Trust in automation: designing for appropriate reliance. *Hum Factors* 2004;**46**:50—80.
9. **Singh IL,** Molloy R, Parasuraman R. Automation induced "complacency": development of the complacency-potential rating scale. *Int J Aviat Psychol* 1993;**3**:111—22.
10. **Cummings ML.** Automation bias in intelligent time critical decision support systems. In: *AIAA 1st Intelligent Systems Technical Conference*. AIAA, 2004:6313. [as referenced in Mckibbon: http://171.67.114.118/content/13/6/653.abstract].
11. **Parasuraman R,** Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors* 2010;**52**:381—410.
12. **Moher D,** Liberati A, Tetzlaff J, *et al*; The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. Published Online First: 21 July 2009;**6**:e1000097. doi:10.1371/journal.pmed1000097.
13. **Wang L,** Jamieson GA, Hollands JG. Selecting methods for the analysis of reliance on automation. *Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors and Ergonomics Society, 2008:287—91.
14. **Landis JR,** Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159—74.
15. **Juni P,** Witschi A, Bloch R, *et al*. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;**282**:1054—60.
16. **Alberdi E,** Povyakalo A, Strigini L, *et al*. Effects of incorrect CAD output on human decision making in mammography. *Acad Radiol* 2004;**11**:909—18.
17. **Tsai TL,** Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. *J Am Med Inform Assoc* 2003;**10**:478—83.
18. **Berner ES,** Maisiak RS, Heudebert GR, *et al*. Clinician performance and prominence of diagnoses displayed by a clinical diagnostic decision support system. *AMIA Annu Symp Proc* 2003:76—80.
19. **Hillson SD,** Connelly DP, Liu Y. The effects of computer-assisted electrocardiographic interpretation on physicians' diagnostic decisions. *Med Decis Making* 1995;**15**:107—12.
20. **Bogun F,** Anh D, Kalahasty G, *et al*. Misdiagnosis of atrial fibrillation and its clinical consequences. *Am J Med* 2004;**117**:636—42.
21. **Dreiseitl S,** Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artif Intell Med* 2005;**33**:25—30.
22. **Guerlain S,** Smith PJ, Obradovich J, *et al*. *The Antibody Identification Assistant (AIDA), an Example of a Cooperative Computer Support System*. Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century. 1995. **2**:1909—14.
23. **Wyatt J.** Lessons learned from the field trial of ACORN, an expert system to advise on chest pain. In: Barber B, Cao D, Qin D, eds. *Proc. Sixth World Conference on Medical Informatics, Singapore.* Amsterdam: North Holland, 1989:111—15.
24. **Westbrook JI,** Coiera EW, Gosling AS. Do online information retrieval systems help experienced clinicians answer clinical questions? *J Am Med Inform Assoc* 2005;**12**:315—21.
25. **McKibbon KA,** Fridsma DB. Effectiveness of clinician-selected electronic information resources for answering primary care physicians' information needs. *J Am Med Inform Assoc* 2006;**13**:653—9.
26. **Southern WN,** Arnsten JH. The effect of erroneous computer interpretation of ECGs on resident decision making. *Med Decis Making* 2009;**29**:372—6.
27. **Sowan AK,** Gaffoor M, Soeken K, *et al*. A comparison of medication administrations errors using CPOE orders vs. handwritten orders for pediatric continuous drug infusions. *AMIA Annu Symp Proc* 2006:1105.
28. **Durieux P,** Nizard R, Ravaud P, *et al*. A clinical decision support system for prevention of venous thromboembolism—effect on physician behaviour. *JAMA* 2000;**283**:2816—21.
29. **Guerlain S.** Factors influencing the cooperative problem-solving of people and computers. *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, 1993:387—91.
30. **Ho G,** Wheatley D, Scialfa CT. Age differences in trust and reliance of a medication management system. *Interact Comput* 2005;**17**:690—710.
31. **Ikeda M,** Ishigaki T, Yamauchi K. A signal-detection experiment measuring the effect of computer-aided detection on radiologists' performance. *Med Decis Making* 2000;**20**:343—51.
32. **Workman M.** Expert decision support system use, disuse, and misuse: a study using the theory of planned behaviour. *Comput Human Behav* 2005;**21**:211—31.
33. **Li F,** Li Q, Engelmann R, *et al*. Improving radiologists' recommendations with computer-aided diagnosis for management of small nodules detected by CT. *Acad Radiol* 2006;**13**:943—50.
34. **Fenton JJ,** Taplin SH, Carney PA, *et al*. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007;**356**:1399—409.
35. **Morimoto T,** Iinuma G, Shiraishi J, *et al*. Computer-aided detection in computed tomography colonography: current status and problems with detection of early colorectal cancer. *Radiat Med* 2008;**26**:261—9.
36. **Zheng B,** Ganott MA, Britton CA, *et al*. Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. *Radiology* 2001;**221**:633—40.
37. **Hadjiski L,** Chan HP, Sahiner B, *et al*. Improvement in radiologists'characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study. *Radiology* 2004;**233**:255—65.
38. **Helvie MA,** Hadjiiski L, Makariou E, *et al*. Sensitivity of non-commercial computer-aided detection system for mammographic breast cancer detection: pilot clinical trial. *Radiology* 2004;**231**:208—14.
39. **Petrick N,** Haider M, Summers RM, *et al*. CT colonography with computer-aided detection as a second reader: observer performance study. *Radiology* 2008;**246**:148—56.
40. **Quek ST,** Thng CH, Khoo JB, *et al*. Radiologists' detection of mammographic abnormalities with and without a computer-aided detection system. *Australas Radiol* 2003;**47**:257—60.
41. **Gakenheimer DC.** The efficacy of a computerized caries detector in intraoral digital radiography. *J Am Dent Assoc* 2002;**133**:883—9.
42. **Moberg K,** Bjurstam N, Wilczek B, *et al*. Computed assisted detection of interval breast cancers. *Eur J Radiol* 2001;**39**:104—10.
43. **Kobayashi T,** Xu XW, MacMahon H, *et al*. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology* 1996;**199**:843—8.
44. **Xu X,** Wickens CD, Rantanen EM. Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information. *Ergonomics* 2007;**50**:112—30.
45. **Skitka LL,** Mosier K, Burdick MD. Accountability and automation bias. *Int J Hum Comput Stud* 2000;**52**:701—17.
46. **Mosier KL,** Skitka LJ, Heers S, *et al*. Automation bias—decision making and performance in high-tech cockpits. *Int J Aviat Psychol* 1997;**8**:47—63.
47. **Bailey NR,** Scerbo MW. *The Effects of Operator Trust, Complacency Potential, and Task Complexity on Monitoring a Highly Reliable Automated System. Dissertation Abstracts International: Section B: The Sciences and Engineering*. US: ProQuest Information & Learning, 2005.

48. **Ma R,** Kaber DB. Situation awareness and driving performance in a simulated navigation task. *Ergonomics* 2007;**50**:1351—64.
49. **Moray N,** Inagaki T, Itoh M. Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *J Exp Psychol Appl* 2000;**6**:44—58.
50. **Wickens C,** Colcombe A. Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Hum Factors* 2007;**49**:5839—50.
51. **Madhavan P,** Wiegmann DA. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Hum Factors* 2007;**49**:5773—85.
52. **Bahner JE,** Huper AD, Manzey D. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *Int J Hum Comput Stud* 2008;**66**:688—99.
53. **Marten K,** Seyfarth T, Auer F, *et al*. Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists. *Eur Radiol* 2004;**14**:1930—8.
54. **Walsham AC,** Roberts HC, Kashani HM, *et al*. The use of computer-aided detection for the assessment of pulmonary arterial filling defects at computed tomographic angiography. *J Comput Assist Tomogr* 2008;**32**:913—18.
55. **Sarter NB,** Schroeder B. Supporting decision making and action selection under time pressure and uncertainty: the case of in-flight icing. *Hum Factors* 2001;**43**:4573—83.
56. **McGuirl JM,** Sarter NB. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Hum Factors* 2006;**48**:4656—65.
57. **Lee JD,** Moray N. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 1992;**35**:1243—70.
58. **Muir BM,** Moray N. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 1996;**39**:429—60.
59. **Prinzel LJ,** Freeman FG, Prinzel HD. Individual differences in complacency and monitoring for automation failures. *Individual Differences Research* 2005;**3**:27—49.
60. **Yeh M,** Wickens CD. Display signaling in augmented reality: effects of cue reliability and image realism on attention allocation and trust calibration. *Hum Factors* 2001;**43**:355—65.
61. **Biros DP,** Daly M, Gunsch G. The influence of task load and automation trust on deception detection. *Group Decis Negot* 2004;**13**:173—89.
62. **Dzindolet MT,** Peterson SA, Pomranky RA, *et al*. The role of trust in automation reliance. *Int J Hum Comput Stud* 2003;**58**:697—718.
63. **Wiegmann DA.** Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Hum Factors* 2002;**44**:44—50.
64. **Dieckmann A,** Dippold K, Dietrich H. Compensatory versus non-compensatory models for predicting consumer preferences. *Judgm Decis Mak* 2009;**4**:200—13.
65. **Burdick MD,** Skitka LJ, Mosier KL, *et al*. The ameliorating effects of accountability on automation bias. *Proceedings of the 3rd Symposium on Human Interaction with Complex Systems*. Dayton, OH: IEEE Computer Society, 1996:142.
66. **Grubb PL,** Warm JS, Dember WN, *et al*. Effects of Multiple-Signal Discrimination on Vigilance Performance and Perceived Workload. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society, 1995:1360—4.
67. **Dixon SR,** Wickens CD. Automation reliability in unmanned aerial vehicle control: a reliance-compliance model of automation dependence in high workload. *Hum Factors* 2006;**48**:474—86.
68. **McFadden SM,** Vimalachandran A, Blackmore E. Factors affecting performance on a target monitoring task employing an automatic tracker. *Ergonomics* 2004;**47**:257—80.
69. **Mosier KL,** Skitka LJ, Dunbar M, *et al*. Aircrews and automation bias: the advantages of teamwork? *Int J Aviat Psychol* 2001;**11**:1—14.
70. **Masalonis AJ.** Effects of training operators on situation-specific automation reliability. *IEEE International Conference on Systems, Man, and Cybernetics*. Washington DC: IEEE Computer Society Press, 2003;**2**:1595—9.
71. **Parasuraman R,** Mouloua M, Molloy R. Effects of adaptive task allocation on monitoring of automated systems. *Hum Factors* 1996;**38**:665—79.
72. **Parasuraman R,** Molloy R, Singh IL. Performance consequences of automation-induced 'complacency'. *Int J Aviat Psychol* 1993;**3**:1—23.
73. **Singh IL,** Molloy R, Parasuraman R. Automation-induced monitoring inefficiency: role of display location. *Int J Hum Comput Stud* 1997;**46**:17—30.
74. **Shiffman RN,** Liaw Y, Brandt CA, *et al*. Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. *J Am Med Inform Assoc* 1999;**6**:104—14.
75. **Campbell EM,** Sittig DF, Guappone KP, *et al*. "Overdependence on technology: an unintended adverse consequence of computerized provider order entry". *AMIA Ann Symp Proc* 2007;**11**:94—8.