Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care

Citation for published version:

Digital Object Identifier (DOI):
10.1177/1460458219900452

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Health Informatics Journal

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care: A systematic review

Kathrin Cresswell, Margaret Callaghan, Sheraz Khan, Zakariya Sheikh, Hajar Mozaffar and Aziz Sheikh
The University of Edinburgh, UK

Abstract
There is growing interest in the potential of artificial intelligence to support decision-making in health and social care settings. There is, however, currently limited evidence of the effectiveness of these systems. The aim of this study was to investigate the effectiveness of artificial intelligence-based computerised decision support systems in health and social care settings. We conducted a systematic literature review to identify relevant randomised controlled trials conducted between 2013 and 2018. We searched the following databases: MEDLINE, EMBASE, CINAHL, PsycINFO, Web of Science, Cochrane Library, ASSIA, Emerald, Health Business Fulltext Elite, ProQuest Public Health, Social Care Online, and grey literature sources. Search terms were conceptualised into three groups: artificial intelligence-related terms, computerised decision support-related terms, and terms relating to health and social care. Terms within groups were combined using the Boolean operator OR, and groups were combined using the Boolean operator AND. Two reviewers independently screened studies against the eligibility criteria and two independent reviewers extracted data on eligible studies onto a customised sheet. We assessed the quality of studies through the Critical Appraisal Skills Programme checklist for randomised controlled trials. We then conducted a narrative synthesis. We identified 68 hits of which five studies satisfied the inclusion criteria. These studies varied substantially in relation to quality, settings, outcomes, and technologies. None of the studies was conducted in social care settings, and three randomised controlled trials showed no difference in patient outcomes. Of these, one investigated the use of Bayesian triage algorithms on forced expiratory volume in 1 second (FEV1) and health-related quality of life in lung transplant patients. Another investigated the effect of image pattern recognition on neonatal development outcomes in pregnant women, and another investigated the effect of the Kalman filter technique for warfarin dosing suggestions on time in therapeutic range. The remaining two randomised controlled trials, investigating computer vision and neural networks on medication adherence

Corresponding author:
Kathrin Cresswell, The University of Edinburgh, Room 115 Medical School, Teviot Place, Edinburgh EH8 9AG, UK.
Email: kathrin.beyer@ed.ac.uk

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
and the impact of learning algorithms on assessment time of patients with gestational diabetes, showed statistically significant and clinically important differences to the control groups receiving standard care. However, these studies tended to be of low quality lacking detailed descriptions of methods and only one study used a double-blind design. Although the evidence of effectiveness of data-driven artificial intelligence to support decision-making in health and social care settings is limited, this work provides important insights on how a meaningful evidence base in this emerging field needs to be developed going forward. It is unlikely that any single overall message surrounding effectiveness will emerge - rather effectiveness of interventions is likely to be context-specific and calls for inclusion of a range of study designs to investigate mechanisms of action.

Keywords
artificial intelligence, decision support systems, narrative synthesis, randomised controlled trial, systematic review

Background
There is an increasing focus on health information technology (HIT) to improve the quality, safety, and efficiency of care. There is also a growing empirical evidence base that knowledge-based computerised decision support (CDS) and, in particular, knowledge-based clinical decision support systems, which form a subcategory of these, have the potential to improve practitioner performance. Such technologies commonly draw on an existing knowledge base of research evidence and/or guidelines to provide logical reasoning-based expert advice. Knowledge-based CDS is different from data-driven CDS in that it does not involve the creation of new knowledge.

Recent reviews have shown that artificial intelligence (AI) algorithms in digital health interventions can be effective in improving health outcomes across a range of conditions, but none has focused on data-driven CDS systems. These systems are designed to emulate human performance typically by analysing large complex data sets. There are now over 16 AI-based products approved by the US Food and Drug Administration (FDA) (https://medium.com/syncedreview/ai-powered-fda-approved-medical-health-projects-a19aba7c681).

There is growing interest from the public, health service providers, policymakers, system vendors, the media, and funding bodies in the potential of CDS linked to data-driven AI-based algorithms, as these can help to quantify risk and facilitate human decision-making. However, there has, to date, been no systematic attempt to scope the empirical evidence base in relation to the effectiveness of CDS systems linked to data-driven AI-based algorithms and some have cautioned against the hype associated with AI-based technologies used in healthcare delivery. A particular gap in evidence relates to social care settings, which has significant opportunities for data-driven CDS (e.g. alerting care workers to potential medical emergencies). In the context of this work, we define social care as health and care–related services associated with assisted living (as opposed to treating illness or injury).

We aimed to investigate the effectiveness of CDS linked to data-driven AI-based algorithms in health and social care settings. Our research questions were as follows:

- What types of CDS linked to data-driven AI are used in health and social care settings?
- What is the evidence surrounding effectiveness of CDS linked to data-driven AI in improving the quality, safety, and efficiency of care?
Methods

Design

We undertook a systematic review of published empirical research. The systematic review protocol is registered with the International Prospective Register of Systematic Reviews (PROSPERO) and reported using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines.8,9

We used the Population-Intervention-Comparison-Outcome (PICO) framework to form the research questions and to focus the literature search (see Table 1).

Search strategy

We searched the published empirical literature from 2013 until September 2018 for work, investigating AI to support decision-making in health and social care settings. We chose this start date as in 2013; IBM’s Watson was first used in the medical field demonstrating the potential usefulness of AI algorithms in healthcare.10

We searched the following databases: MEDLINE, EMBASE, CINAHL, PsycINFO, Web of Science, Cochrane Library, ASSIA, Emerald, Health Business Fulltext Elite, ProQuest Public Health, Social Care Online, and grey literature sources. These were chosen as they provided disciplinary breadth covering clinical and social care as well as various outcomes, including clinical and population health. Grey literature sources were included to ensure that we did not miss any potentially important studies that, for example, may not have been published because of negative findings.

We divided search terms into three groups: AI-related terms, CDS-related terms, and terms relating to health and social care settings. Terms within groups were combined using the Boolean operator OR, and groups were combined using the Boolean operator AND. We applied methodological filters to find randomised controlled trials (RCTs) and provide search strategies for each database in Online Appendix 1. Focusing on RCTs allowed us to assess the effectiveness of CDS linked to data-driven AI-based algorithms using studies at the lowest risk of bias, which is important, given the current hype of AI-based technologies and the need for rigorous testing of applications.

Eligibility criteria

Studies were eligible for inclusion if they were conducted in health and social care settings and published in English, if they focused on data-driven AI, and if they used technological systems for clinical, managerial, and self-management decision-making.

Table 1. PICO framework for the review.

<table>
<thead>
<tr>
<th>P (Population)</th>
<th>The intended users of data-driven AI to support decision-making, including health and social care users (patients and citizens), health and social care professionals, and managers</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (Intervention)</td>
<td>Data-driven AI-based CDS systems</td>
</tr>
<tr>
<td>C (Comparator(s))</td>
<td>Non-AI-based approaches</td>
</tr>
<tr>
<td>O (Outcomes)</td>
<td>Practitioner performance (as this has been shown to be improved in previous reviews of the effectiveness of CDS)</td>
</tr>
<tr>
<td></td>
<td>Patient, citizen, and population outcomes (as improved quality, safety and efficiency are key aims of HIT implementation)</td>
</tr>
<tr>
<td></td>
<td>Health system outcomes (as improved quality, safety and efficiency are key aims of HIT implementation)</td>
</tr>
</tbody>
</table>

AI: artificial intelligence; CDS: computerised decision support; HIT: health information technology.
We excluded studies if they were not RCTs or if they fell outside our scope of interest. This included, for example, studies that evaluated technology that is not commonly associated with systems driven by the analysis of patterns and models emerging from very large data sets, and those that did not focus on a combination of CDS and AI.

**Study selection**

Titles and abstracts of studies identified from the searches were screened by two investigators (M.C. and S.K.), who screened all retrieved potentially eligible studies independently against the above criteria. Any disagreements were resolved by discussion or, if necessary, through arbitration by K.C.

**Quality assessment and analysis**

Formal quality assessment of eligible studies was undertaken independently by two reviewers (M.C. and Z.S.) using the Critical Appraisal Skills Programme (CASP) checklist for RCTs.\textsuperscript{11} Disagreements were resolved by discussion or, if necessary, through arbitration by K.C.

**Data extraction**

M.C. and Z.S. extracted data onto a customised data extraction sheet in Microsoft Excel. This comprised the following headings: authors, title, journal, year, country, healthcare setting, participant number and type, age, timescale, type of AI, type of decision support, comparator (non-AI/CDS-based approaches), health problem/condition, outcomes assessed, impact on practitioner performance, impact on patient outcomes, impact on patient self-management, other estimates of effectiveness, enablers and barriers, reviewer notes, and reviewer interpretation. We included specific outcomes associated with CDS-based approaches in the existing empirical evidence base, so that we could ascertain where improvements occurred.

**Data analysis**

A meta-analysis was inappropriate due to the heterogeneity of technologies and care contexts. Data were therefore descriptively summarised and synthesised. There is a lack of guidance surrounding the narrative synthesis of quantitative data, but we have followed steps outlined by Popay et al:\textsuperscript{12} (1) developing a preliminary synthesis describing the various functions of technological systems and the study contexts, (2) summarising evidence of effectiveness and study quality, and (3) exploring relationships in the data.

**Results**

We identified 69 potentially eligible studies resulting from our searches. After removing duplicates, we screened 68 abstracts. At screening stage, reviewers dropped 31 abstracts. Most excluded abstracts \((n=16)\) did not include AI. Study protocols \((n=10)\) and non-randomised studies \((n=10)\) were also excluded. We assessed 37 full-text articles for eligibility, of which we excluded 32. Ineligible articles that did not combine AI and CDS functionality \((n=15)\), did not have AI as their primary focus \((n=7)\), did not have RCT designs \((n=6)\), did not report our outcomes of interest \((n=2)\), or did not have interventions that were data-driven \((n=2)\) were excluded.
Five articles were included in the final review. Of these, two were conducted in the United States, one in Spain, one in Denmark, and one in the United Kingdom. Two RCTs took place in inpatient settings, two in a home care setting, and one in an outpatient setting. Key characteristics of included studies are summarised in Table 2.

**Descriptive synthesis: functions of technological systems and study contexts**

We observed substantial variations in the size of studies (from 28 to 47,062 participants), technological systems (i.e. type of AI-based CDS), and timescales over which the effectiveness of systems was assessed. All included studies focused on specific patient populations often with long-term conditions. These included women with gestational diabetes, women in labour, adults with ischaemic stroke, thrombosis patients, and lung transplant recipients.

Types of AI-facilitated decision support also varied widely across studies. One examined learning algorithms to support patient self-management, while another assessed algorithms facilitating automated triaging based on existing data sets. A third study assessed algorithms facilitating the interpretation of foetal cardiotocographs (CTGs) through image pattern recognition. The final two studies assessed the use of neural networks to facilitate medication adherence and the Kalman filter technique (an algorithm using temporal measurements) to personalise warfarin dosing recommendations for patient self-management.

Three studies investigated decision-making in patients, whereas the others focused on decision-making in healthcare professionals. We further found large variations in timescales of studies from 12 weeks to 3.5 years in duration.

Studies most frequently assessed the impact of the intervention on patient outcomes, but these varied significantly across RCTs due to the differences in study populations. For example, Finkelstein et al. assessed forced expiratory volume in 1 second (FEV1) and health-related quality of life in lung transplant patients, while others assessed neonatal development outcomes, medication adherence, and time in therapeutic warfarin range. Two studies also assessed impacts on practitioner performance. These RCTs examined the effectiveness of triaging interventions in lung transplant recipients and the impact on clinician assessment time.

Caballero-Ruiz et al. applied a learning algorithm to a CDS to manage treatment of patients with gestational diabetes through telemedicine. They assessed access to specialised healthcare, evaluation time for patients, adverse gestational diabetes outcomes, clinical time required per patient, number of face-to-face visits, frequency and duration of telematic reviews, patient compliance, and patient satisfaction.

Finkelstein et al. investigated the effectiveness of a Bayesian triage algorithm for automated triaging based on analysing data from a home monitoring programme in lung transplant patients and assessed the effectiveness of triaging clinical interventions compared with manual nurse decision.

The INFANT Collaborative Group tested the effectiveness of image pattern recognition in the interpretation of CTGs of women in labour and assessed neonatal outcomes of development at 2 years of age compared to usual care.

A study conducted by Labovitz et al. assessed the effectiveness of computer vision and neural networks in improving medication adherence in patients with ischaemic stroke and compared it with no daily monitoring.

The final study conducted by Nielsen et al. assessed the effectiveness of an algorithm on warfarin dosing recommendations to patients to prevent thromboembolic events. The control arm included no AI algorithm-based support, with the dosage suggestion equalling the previous week’s dose of warfarin.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Country</th>
<th>Setting</th>
<th>Participants</th>
<th>Timescale</th>
<th>Type of AI</th>
<th>Comparator</th>
<th>Main findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caballero-Ruiz et al.¹³</td>
<td>Spain</td>
<td>Home health setting</td>
<td>450 pregnant women</td>
<td>17 months</td>
<td>Learning algorithm to manage the treatment of patients with gestational diabetes through telemedicine</td>
<td>Standard care</td>
<td>Assessment time decreased by almost a third. Face-to-face time was reduced by 88% but overall time remained the same. Automatic detection of 100% of patients who needed insulin therapy and diet adjustment. High patient satisfaction.</td>
</tr>
<tr>
<td>Finkelstein et al.¹⁴</td>
<td>USA</td>
<td>Inpatient setting</td>
<td>65 lung transplant recipients between the age of 37 and 69 years</td>
<td>October 2006 to April 2009 plus 1-year follow-up</td>
<td>Computer based Bayesian triage algorithm for automated triaging</td>
<td>Manual nurse decision</td>
<td>No differences in outcomes measured.</td>
</tr>
<tr>
<td>INFANT Collaborative group¹⁵</td>
<td>UK and Ireland</td>
<td>Labour ward</td>
<td>Women in labour &gt;35-week gestation, 16 years of age or older</td>
<td>January 2010 to August 2013</td>
<td>Image pattern recognition in the interpretation of CTGs</td>
<td>Standard care</td>
<td>No difference in any neonatal outcome of development assessment at 2 years of age.</td>
</tr>
<tr>
<td>Labovitz et al.¹⁶</td>
<td>USA</td>
<td>Outpatient care</td>
<td>28 adults with recently diagnosed ischaemic stroke receiving anticoagulation therapy</td>
<td>12 weeks</td>
<td>Computer vision and neural networks</td>
<td>No daily monitoring</td>
<td>Mean cumulative adherence based on the AI platform was 90.5%. Mean cumulative adherence indicated by plasma drug concentration was 100% for the intervention and 50% for the control group. Mean cumulative adherence indicated by pill count was 97.2% for the intervention and 90.6% for the control group. For patients receiving direct-acting oral anticoagulants, mean cumulative adherence based on the AI platform was 90.1%, mean cumulative adherence indicated by pill count was 96.4% for the intervention and 90.9% for the control group, and mean cumulative adherence indicated by plasma drug concentration was 100% for the intervention and 33% for the control group.</td>
</tr>
<tr>
<td>Nielsen et al.¹⁷</td>
<td>Denmark</td>
<td>Home health setting</td>
<td>191 patients with an indication for warfarin treatment who were in steady-state patient self-management treatment with a mean age of 65 years</td>
<td>September 2014 to November 2014, followed-up for at least 90 days with a mean follow-up of 140 days</td>
<td>Personalised support for warfarin dosing (Kalman filter technique)</td>
<td>No AI algorithm-based support (the dosage suggestion in the placebo arm would equal last week’s dose of warfarin)</td>
<td>The intervention arm achieved a time in therapeutic (INR, international normalised ratio) range (TTR) of 81.6, while the placebo arm attained a TTR of 80.9 (difference (intervention arm minus placebo arm): 0.67 (95% confidence interval)).</td>
</tr>
</tbody>
</table>

AI: artificial intelligence; CTG: cardiotocograph; TTR: time in therapeutic range.
Study quality and evidence of effectiveness

The quality of studies was extremely variable. Details of methods were in some instances difficult to find and only one study used a double-blind design. In another study, only patients were blinded, and another one used no blinding. For the remaining two, it was unclear whether blinding took place (Table 3).

We observed mixed evidence of effectiveness, with two studies showing no statistically significant difference to the control group, and three showing statistically significant and clinically relevant differences between the intervention and the control groups. Detailed study characteristics and outcomes are provided in Table 2.

One study with high risk of bias, focussing on a learning algorithm to help with managing gestational diabetes, reported positive findings. This RCT showed a decrease in assessment time (from 15 minutes in the control group to 2.778 ± 0.858 minutes in the intervention group per patient) and a reduction in face-to-face consultations (3.207 ± 2.846 visits in the control group and 0.367 ± 0.901 in the intervention group). Another high risk of bias study using computer vision and neural networks reported the mean cumulative medication adherence indicated by plasma drug concentration to be 100 per cent for the intervention and 33 per cent for the control group.

Other studies with low risk of bias showed no difference. One trial used a Bayesian algorithm tool for remote monitoring, follow-up, and triage of patients after lung transplants. The authors reported no difference in the detection of changes in patients’ FEV1 and quality of life between intervention and control groups. Both groups showed non-significantly different decreases over two years, including a two per cent FEV1 decrease (p = 0.721) at Year 1 and a three per cent decrease at Year 2 (p = 0.861).

Another trial drawing on image pattern recognition for the computerised interpretation of CTGs during labour did not show an effect on neonatal outcomes. Poor neonatal outcomes were reported in 172 (0.7%) babies in the AI-based CDS group versus 171 (0.7%) in the control group.

A third trial showed no difference when comparing personalised algorithm–generated warfarin dosing recommendations for thrombosis patients with standard care and showed that the intervention achieved a time in therapeutic range of 81.6, while the control group achieved 80.9 (difference: 0.67 (95% confidence interval −2.93 to 4.27)).

Exploring relationships in the data

All three studies that showed statistically significant differences between the intervention and the control groups were conducted in a home/outpatient setting of patients with chronic conditions. Here, all CDS interventions targeted medications (warfarin, insulin, and anticoagulation). The
studies that did not show statistically significant differences were conducted in acute care settings and CDS interventions here targeted diagnostic interventions, including interpretation of CTGs and nurse triaging. It is, therefore, possible that the lack of statistical significance is associated with diagnostic uncertainty.

Discussion

Strengths and limitations

Our review is a first of type examining the use of data-driven AI to support decision-making. However, as we have shown, the number of potentially relevant RCTs is limited, perhaps reflecting the immaturity of the field (10 of the excluded studies were protocols), but also potentially due to overlapping definitions surrounding CDS and AI (23 of the excluded studies were not classified as AI, and 15 did not combine AI and CDS). For example, it was at times hard for the research team to distinguish between knowledge-driven and data-driven applications. Possible reasons for the paucity of RCTs may include ethical and legal considerations around data access and medical device regulations, the long and costly nature of RCTs, and the changing (often self-improving nature) of AI tools. Nevertheless, although the review may have been conducted early, there is, at present, no guidance on how to develop the existing evidence base around this important, but still a developing topic. There are therefore some important recommendations for future work emerging from our review.

Future directions for research emerging from this work

To develop a meaningful evidence base, designs that investigate the hypothesised mechanisms of action associated with specific technological functionalities, spanning data-driven AI, and CDS may be an appropriate first step to establishing the effectiveness of such complex interventions. These may include qualitative evaluations, either as stand-alone studies or embedded with RCTs to investigate mechanisms of action. There is also a need to look at potential unintended consequences and challenges associated with novel systems in combination with RCTs. Concurrent qualitative evaluation can help to address some of these issues and also help to identify contextual dynamics and potential reasons for effectiveness (or lack thereof).

Quantitative work needs to carefully conceptualise control groups. For example, all included studies compared AI-based CDS with standard care. In future work, the comparison should tackle AI-enabled CDS versus CDS to see whether AI makes a difference to standard knowledge-based CDS. It may be that the lack of existing RCTs in the area is due to issues with data access for AI specialists. This may also help to explain the involvement of system developers in the majority of our included studies – they may have had privileged access to data in their systems. The more data algorithms can draw on, the more effective they become, but access to large curated data sets on which researchers can train algorithms is currently still hard to achieve. The issues with quality in these studies could also reflect different standards of evidence in computer science and the biomedical sciences.

More generally, there is a need to remember that, as in knowledge-driven CDS, the ultimate responsibility of the decision still lies with the human. As such, those at the receiving end of data-driven AI-based CDS need to be trained to make decisions informed by these systems. This may require developing new skills and/or ways of considering evidence.

Policy recommendations and implications for practice emerging from this work

Our work may support those cautioning against the assumed effectiveness of AI and the associated hype surrounding these technologies. Policymakers need to be aware that evidence
of effectiveness is limited at this stage. To address the variability of existing work in this area, strategic decision makers may need to extract key areas of focus for research and innovation within their locales where applications have the greatest potential to meet a major service need and where they are most likely to deliver real impact. Ideally, work should aim to be comparable in terms of technologies and disease areas and include qualitative formative evaluation components to capture emerging challenges. The limited details reported in the method sections of included studies, particularly in relation to AI algorithms, also calls for clearer standards of reporting of studies to ensure rigour and independent assessment of risk of bias.

As the application of AI is gaining momentum, there is likely to be an increasing need for developing associated evaluation frameworks, reporting guidelines and understanding transferability beyond experimental contexts. A focus on unintended consequences, positive or negative, should be fundamental to these efforts.

**Conclusion**

The findings of this review may lead to the conclusion that the evidence of effectiveness of AI to support decision-making in health and social care settings is limited and contradictory, perhaps reflecting the variability in contexts in which technologies are deployed. We found a very small number of relevant studies with large variability in quality, settings, outcomes, and technologies. We identified no studies from social care settings and none included work investigating any enablers and/or barriers for the use of data-driven AI to support decisions. Three RCTs showed no difference, whereas two of the trials included in this work showed substantial gains, but there are concerns in relation to the quality of these studies.

While the applied review methodology may be viewed as immature in an area where robust empirical evidence is nascent, it has helped to identify how the existing empirical evidence base should be conceptualised in this important but emerging field going forward. There is, therefore, now a need to build on this work to judge the effectiveness of the emerging field of data-driven AI-based CDS.

**Acknowledgements**

The authors thank Healthcare Improvement Scotland for helping with conducting the searches, and Dr Ann Wales for her advice throughout the work. We would also like to thank two anonymous peer reviewers for their helpful comments on an earlier version of this manuscript.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article has drawn on a programme of independent research funded by the Digital Health and Care Institute and Scottish Government. The views expressed are those of the author(s) and not necessarily those of the funders.

**Supplemental material**

Supplemental material for this article is available online.

**ORCID iD**

Kathrin Cresswell https://orcid.org/0000-0001-6634-9537
References