BMJ Health & Care Informatics

Applying a user-centred design machine learning toolkit to an autism spectrum disorder use case

Joseph M Plasek [1,2] Li Zhou [1,2]

To cite: Plasek JM, Zhou L. Applying a user-centred design machine learning toolkit to an autism spectrum disorder use case. *BMJ Health Care Inform* 2023;**30**:e100765. doi:10.1136/bmjhci-2023-100765

Received 16 March 2023
Accepted 15 April 2023

[1]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Somerville, Massachusetts, USA
[2]Medicine, Harvard Medical School, Boston, MA, USA

Correspondence to
Dr Li Zhou;
lzhou@bwh.harvard.edu

Two *BMJ Health & Care Informatics* editors' choice papers present insights based on case studies from real-world data and machine learning models for clinical risk prediction use cases. Seneviratne *et al* focus on case management to demonstrate how one might implement their proposed user-centred design toolkit consisting of process maps, storyboards and four questions.[1] This toolkit was developed to address the tendency to develop machine learning models in an opportunistic manner based on the availability of data rather than through fundamental design principles focused on resolving the actual pain points of stakeholders. Chen *et al* created a screening tool for the early detection of patients with autism spectrum disorder (ASD).[2] Here, we present a critical thought exercise using the four questions in the user-centred design toolkit, applied to the ASD screening tool. Any gaps identified are not intended to serve as criticism of the ASD tool, but serve as illustrative examples of the potential utility of the user-centred design toolkit. The authors examine the ASD tool from the perspectives of a conversation between clinicians and developers.

Question 1: Where are the current pain points? From the clinical perspective, early identification of patients with ASD is critical for active brain development, which is influenced by both genetics and experience. While benefits from early intensive applied behaviour analysis can reduce the need for support services such as occupational, physical and speech therapy over their lifetimes, current ASD screening tools perform poorly.[2–4] From an artificial intelligence perspective, we often have a hard time identifying ASD aetiology. Most cases are idiopathic, but sometimes a cause is known (eg, fragile X syndrome). A structural causal model (often depicted as a directed acyclic graph) can be helpful to explore potential relationships between model features. For example, co-occurrences of ASD with anxiety and attention-deficit/hyperactivity disorder are common, as well as other neurological disorders, and recent studies have shown a link between ASD and the gut microbiome, but correlations are not necessarily causative.[5] More concerning, individual confounders, such as vaccination and circumcision, have repeatedly been shown not to cause ASD but are circumstantial; as ASD unfolds over the first few years of life, a child can appear typical in the first few months before regressing during the time when babies are being vaccinated.[6]

Question 2: Where could machine learning add unique value? Large-scale retrospective clinical claims data contains potentially meaningful causal signals among spurious correlations that can be identified through application of various data science techniques. Chen *et al* used Lasso regularisation and random forests for dimensionality reduction in order to identify complex and hidden correlations in the data. Machine learning may be helpful in addressing temporal biases in the data. For example, prior research conducted over the same time frame of Chen *et al* suggests an artificial increased prevalence due to improved awareness and changes in the diagnostic criteria of ASD.[7] In Chen *et al*, the key predictor of ASD was male, however, increased clinical attention to identifying ASD in females over the years of the study has demonstrated a statistical decrease in the male/female ratio which may manifest as drift in machine learning models.[2 7]

Question 3: How will the model output be acted on? The authors do not propose a specific clinical workflow within which this ASD screening tool could be used, though they speculate that it could be used as a triaging tool for identifying patients that would benefit from a comprehensive diagnostic evaluation that involves the integration

of behavioural symptoms in the context of developmental history, family factors and cognitive level.

Question 4: What criteria should the model be optimised for? Screening tests are helpful if they can rule-in (Specific test when Positive rules IN the disease) or rule-out (Sensitive test when Negative rules OUT the disease) a diagnosis.[8 9] Given that ASD is a rare (2.3%), underdiagnosed disorder, choosing a threshold based on the positive predictive value (PPV)-sensitivity curve (also known as thearea under the Precision-Recall curve (AUPRC)) may be challenging and suboptimal because PPV, in this case, is severely adversely affected by prevalence. Chen *et al* report PPV at three sensitivity target thresholds, with minimal improvements to PPV. Instead, optimising and choosing thresholds based off of the often-forgotten negative predictive value (NPV)-specificity curve may be more appropriate as early ASD interventions are costly, and having a high NPV reassures providers that the child is not likely being harmed through inaction or delay.[3 4]

This thought exercise adds insights to the existing evaluation of the ASD screening tool. This suggests that the Seneviratne *et al* toolkit is a potentially useful practical addition to the multitude of clinical machine learning guidelines, emphasising the utility of connecting with stakeholders to codesign models that are clinically meaningful and implementable in real-world workflows.[1]

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iDs
Joseph M Plasek http://orcid.org/0000-0002-9686-3876
Li Zhou http://orcid.org/0000-0003-3874-4833

REFERENCES

1 Seneviratne MG, Li RC, Schreier M, *et al*. User-centred design for machine learning in health care: a case study from care management. *BMJ Health Care Inform* 2022;29:e100656.
2 Chen Y-H, Chen Q, Kong L, *et al*. Early detection of autism spectrum disorder in young children with machine learning using medical claims data. *BMJ Health Care Inform* 2022;29:e10054.
3 Sampaio F, Feldman I, Lavelle TA, *et al*. The cost-effectiveness of treatments for attention deficit-hyperactivity disorder and autism spectrum disorder in children and adolescents: a systematic review. *Eur Child Adolesc Psychiatry* 2022;31:1655–70.
4 Rodgers M, Marshall D, Simmonds M, *et al*. Interventions based on early intensive applied behaviour analysis for autistic children: a systematic review and cost-effectiveness analysis. *Health Technol Assess* 2020;24:1–306.
5 Taniya MA, Chung H-J, Al Mamun A, *et al*. Role of gut microbiome in autism spectrum disorder and its therapeutic regulation. *Front Cell Infect Microbiol* 2022;12:915701.
6 Farrington CP, Miller E, Taylor B. MMR and autism: further evidence against a causal association. *Vaccine* 2001;19:3632–5.
7 Solmi M, Song M, Yon DK, *et al*. Incidence, prevalence, and global burden of autism spectrum disorder from 1990 to 2019 across 204 countries. *Mol Psychiatry* 2022;27:4172–80.
8 Umberger RA, Hatfield LA, Speck PM. Understanding negative predictive value of diagnostic tests used in clinical practice. *Dimens Crit Care Nurs* 2017;36:22–9.
9 Baeyens J-P, Serrien B, Goossens M, *et al*. Questioning the "spin and snout" rule in clinical testing. *Arch Physiother* 2019;9:4.

BMJ Health & Care Informatics

Willingness of diabetes mellitus patients to use mHealth applications and its associated factors for self-care management in a low-income country: an input for digital health implementation

Agmasie Damtew Walle ,[1] Tigist Andargie Ferede,[2] Adamu Ambachew Shibabaw ,[1] Sisay Maru Wubante ,[3] Habtamu Alganeh Guadie,[4] Chalachew Msganaw Yehula ,[3] Addisalem Workie Demsash [1]

To cite: Walle AD, Ferede TA, Shibabaw AA, *et al*. Willingness of diabetes mellitus patients to use mHealth applications and its associated factors for self-care management in a low-income country: an input for digital health implementation. *BMJ Health Care Inform* 2023;**30**:e100761. doi:10.1136/bmjhci-2023-100761

Received 09 March 2023
Accepted 20 April 2023

For numbered affiliations see end of article.

Correspondence to
Mr Agmasie Damtew Walle; agmasie89@gmail.com

ABSTRACT

Background Although mHealth applications are becoming more widely available and used, there is no evidence about why people are willing to use them. Therefore, this study aimed to assess the willingness of patients with diabetes to use mHealth applications and associated factors for self-care management in Ethiopia.

Methods An institutional cross-sectional study was conducted among 422 patients with diabetes. Data were collected using pretested interviewer-administered questionnaire. Epi Data V.4.6 for entering the data and STATA V.14 for analysing the data were used. A multivariable logistic regression analysis was carried out to identify factors associated with patient's willingness to use mobile health applications.

Results A total of 398 study participants were included in the study. About 284 (71.4%) 95% CI (66.8% to 75.9%)). Of participants were willing to use mobile health applications. Patients below 30 years of age (adjusted OR, AOR 2.21; 95% CI (1.22 to 4.10)), urban residents (AOR 2.12; 95% CI (1.12 to 3.98)), internet access (AOR 3.91; 95% CI (1.31 to 11.5)), favourable attitude (AOR 5.20; 95% CI (2.60 to 10.40)), perceived ease of use (AOR 2.57; 95% CI (1.34 to 4.85)) and perceived usefulness (AOR 4.67; 95% CI (1.95 to 5.77)) were significantly associated with patients' willingness to use mobile health applications.

Conclusions Overall, diabetes patients' willingness to use mobile health applications was high. Patients' age, place of residence, internet access, attitude, perceived ease of use and perceived usefulness were significant factors concerning their willingness to use mobile health applications. Considering these factors could provide insight for developing and adopting diabetes management applications on mobile devices in Ethiopia.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Ethiopia faces obstacles in adopting a sustainable mobile health application due to a lack of top-level commitment to using the system for managing chronic disease.
⇒ Mobile health application allows users to store their health-related data, provide medical references and support clinical decision-making.
⇒ For the adoption of mHealth interventions, considering users' willingness is crucial.

WHAT THIS STUDY ADDS

⇒ This study assessed the willingness of patients with diabetes mellitus (DM) to use mhealth apps in Ethiopia, which aided in the development of mobile health technology for the advancement of Ethiopia's healthcare system.
⇒ The results of this study were used as input to design and test the effectiveness of a mobile health application for reducing the burden of DM in Ethiopia.
⇒ The study provides potential solutions for the identified barriers to the willingness of using mobile health applications.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ There is limited evidence on the willingness to use mobile health applications to manage diabetes and it serves as a baseline for researchers in a resource-limited setting.
⇒ Practically, this study offers insights for policymakers, developers, managers and decision-makers in the healthcare industry to improve the use and willingness of mobile health applications for self-care management based on the findings.
⇒ The study serves as a foundation for more interventional research that can create and evaluate mobile health applications as interventions and as a tool for enhancing Ethiopia's DM prevention programme.

INTRODUCTION

According to the International Diabetes Federation 2019, diabetes mellitus (DM) is a

serious public health issue that affects 463 million people worldwide as of 2019 and is expected to affect 578 million people by 2030 and 700 million people by 2045.[1] It is also a prevalent, costly, chronic metabolic condition that is defined by elevated blood glucose levels brought on by either an inability to produce insulin (type 1 diabetes) or an inability to produce enough insulin and insulin resistance (type 2 diabetes).[1 2] Despite the WHO's attempts to reduce the strain of diabetes, its prevalence is rising and might lead to more early deaths and an estimated US$2.1 trillion (2.2% of the global gross domestic product) in economic impacts by 2030.[3]

Ethiopia is 1 of the 48 nations in the International Diabetes Federation's African region (AFR), and currently, 24 million individuals in the AFR region and 537 million people worldwide have diabetes; by 2045, those numbers will rise to 55 million, and 3.3% of adults have the disease.[4] Diabetes can lead to a variety of consequences, including retinopathy, hypertension, cardiovascular, nephropathy and macrovascular disorders, all of which lower patient quality of life, reduce rates of economic growth, reduce labour productivity and raise healthcare costs. It is critical to achieve optimal glycaemic management to avoid and reduce problems.[5] Consequently, digital health solutions, such as mobile health technologies, are essential for overcoming time and geographical constraints through mobile applications, and remote monitoring of data at home, such as blood glucose levels.[6]

Mobile health applications use the internet to support medical and health activities, offer tools for tracking consumers' health states, storing their health-related data, providing medical references and supporting clinical decision-making.[7] A self-contained programme or piece of software with a specific purpose is referred to as an 'application' or 'app' and is typically customised to run on mobile devices, including smartphones, tablets and wearable technology.[8]

Consumers are using mobile health applications more frequently as a result of the global proliferation of mobile device technology.[9] More than 2.5 billion individuals will own smartphones by the year 2019, and by 2017, more than 50% of them will have mobile health applications installed.[10] More than 60% of people in the USA use digital devices and mobile health apps to manage their health.[11] In China, the most popular mobile health applications had over 10.5 million active users as of January 2020.[12] In a study done in Japan, although only 51 (16%) people currently use information communication technology (ICT)-based self-management tools, 157 people (50%) said they would be willing to use them.[5] Patients with diabetes have access to mobile applications that have increased their physical activity and hypoglycaemia control.[5 13] As a result, mobile health applications are becoming a crucial part of managing personal health.

To use mHealth technologies for managing and caring for diabetics, it is critical to assess the level of the patient's willingness and identify factors for using mobile health applications. Accordingly, factors were identified, such as sociodemographic factors (age, gender, educational status, place of residence).[1 5 6 8 14–17] The technology acceptance model; and the theory of reasoned action, state the adoption of new technology is dependent on the user's willingness or intention, which is influenced by attitude, perceived ease of use and perceived usefulness.[14 16 17]

The findings may have implications for practice, policy and upcoming researchers. The main beneficiaries of this study, which is useful as input for common practices, are the patients, health professionals, regional health bureau and non-government organisations. According to our review of the literature, there is no evidence of research exploring how willing Ethiopian patients with diabetes are to use mobile health applications to manage their health. This study aimed to assess the willingness of patients with DM to use a mobile health application and its associated factors in southwest Ethiopia.

METHODS

Study area and period

The study was carried out in public facilities in Ilu Abba Bor and Buno Bedelle Zones, Oromia Regional State, southwest Ethiopia. Ilu Abba Bor Zone and Buno Bedelle Zone are two of the Oromia regional state situated southwest of the region and located at a distance of about 600 km and 483 km from the centre of the region, respectively. In the two zones, there are five public hospitals, namely: Bedele hospital, Darimu hospital, Dembi hospital, Metu Karl hospital and Chora hospital. The study was conducted from 12 November 2022 to 21 December 2022.

Study design

An institution-based cross-sectional study was carried out among patients with diabetes who were followed up in public hospitals.

Study population

All adult patients with DM who attend public hospitals in the Ilu AbaBor and Bunno Bedelle zones are used as the source population. All adult (>18-year-old) patients with DM at public hospitals in the Ilu AbaBor and Bunno Bedelle zones were included in the study, whereas patients who are seriously ill and unable to give a response during the study period were not included.

Study variables and measurements

The outcome variable was the willingness to use a mobile health application and its associated factors for self-care management. Perceived usefulness, perceived ease of use, attitude, sociodemographic characteristics, clinical and behavioural attributes, and mobile device utilisation patterns were considered as predictor variables for the outcome of interest in this study.

Willingness to use a mHealth app was defined as the user's likelihood to use a computer program or software health application designed to run on a mobile device for self-care management to measure willingness,

respondents were asked whether they would be willing to use mobile phone-based diabetic health applications. Thus, it was measured by the median score. If the score was above the median, it was considered as willing to use mobile health applications, else not willing to use it. Whereas, items for the composite variables were scored on a Likert-type scale with a maximum score of 5 and a range of 1 for 'strongly disagree' to 5 for 'strongly agree'. To produce a composite variable scale (ranging from score 1 to 5) for data analysis, item scores for each composite variable were added and divided by the number of items. Finally, depending on the final result, the composite variable score was dichotomised as 'yes' or 'no'. Accordingly, final scores of 3 or below (strongly disagree, disagree and neutral) were classified as '"no', while final scores of three or above (agree and strongly agree) were classified as 'yes'.[16 18]

Sample size determination and sampling procedure

The sample size was determined using the single population proportion formula by the following assumptions.

$n = \frac{(Z\alpha/2)2\times p(1-p)}{d^2}, = \frac{(1.96)2\times 0.5(0.5)}{(0.05)^2}, = 384$, then after we consider non-response rate 10%.

Finally, 384+384 (0.1)=422, where n=estimated sample size; p=single population proportion (50%); because willingness of patients with DM to use mobile applications in Ethiopia was not investigated, Z/2=95% level of CI; d2=5% margin of error. A total of 422 patients with diabetes who participated in this study were recruited using a systematic random sampling technique. To select the participants, first, the entire sample size was proportionally allocated according to the number of patients with DM in each hospital, resulting in 1874 eligible participants in the study. Then, a sampling interval of four was computed after the total sample size was determined. As a result, every fourth patient who visited the specified hospitals for diabetes follow-up included in this study.

Data collection tool, data quality control and procedure

Data were gathered using standardised, pretested, interviewer-administered questionnaires that were adapted from available research.[1 5 6 8 14–17] The questionnaire contains sociodemographic characteristics, clinical and behavioural characteristics, mobile pattern utilisation and perceived ease of use, perceived usefulness, attitude and willingness to use mHealth applications, which were adapted from Davis's study.[19] For ease of data collection, an English version of the questionnaire was created and translated into an Amharic version. The Amharic version was translated back into English by a language expert to ensure that the meaning was consistent.

To evaluate the validity and reliability of the data collection instrument before the actual data collection, a pretest study outside of the study area was conducted in Jimma Hospital with 10% of the total sample size, and necessary modifications were made accordingly. Cronbach's alpha was used to evaluate the internal consistency for each aspect of the data collection instrument, and it scored on attitude (Cronbach's alpha=0.78), perceived usefulness (Cronbach's alpha=0.82), perceived ease of use (Cronbach's alpha=0.87) and willingness to use mHealth applications (Cronbach's alpha=0.91). Finally, for the actual data collection, 2 days of training were provided for four nurses, two health informatics professionals who were data collectors and three onsite supervisors.

Data processing and analysis

Data entry was done by using Epi Data V.4.6, and analysis was done by using STATA V.14. For descriptive statistics, frequencies and percentages were determined and presented using graphs and tables. A candidate for multivariable, binary logistic regression analysis with a p<0.2 at 95% CI was applied using bivariable binary logistic regression. After that, using the backward technique to enter candidate variables into a multivariable logistic regression model, it was possible to adjust for potential confounders and determine which components were statistically significant for the outcome variable. The adjusted OR (AOR) and its 95% CIs, p<0.05 were used to summarise the findings.

The variance inflation factor was used with a cut-off point of 10 to determine whether multicollinearity existed among independent variables, and there was no evidence of it. Finally, the model fit was examined using the Hosmer and Lemeshow goodness of fit test, and the result showed good goodness of fit for the data (p=0.706).

RESULTS

Sociodemographic characteristics of patients with DM

A total of 398 study participants were included in the study, with a response rate of 94.3%. Of the 398 respondents, 224 (56.3%) were male, 155 (38.1%) were between the ages of 30 and 45, and the mean age was 43+14.6 years. More than half of the 276 respondents (69.3%) lived in urban areas. In addition, 249 (62.6%) of the respondents were married (table 1).

Clinical and behavioural characteristics of the patients

The majority of patients, 249 (62.6%), had type 2 DM. Almost half of the participants 193 (48.5%), had comorbidities. A total of 175 (44%) of these patients were diagnosed 3 years ago or more. A total of 108 (27.1%) of patients were educated by healthcare providers during the follow-up period. Seventy-five (21.4%) of patients had used substances, 156 (39.2%) had a weekly habit of physical exercise, 302 (75.9%) of patients obtained diabetes medication regularly and 113 (28.4%) of patients had a habit of excessive sugar consumption (table 2).

Mobile device utilisation pattern

About three-fourth, 298 (74.9%) of patients with DM had a mobile device, and 216 (57.2%) were using a smart mobile phone. Fifty-two (13.1%) of respondents used mobile applications for disease management, and 271 (68.1%) of them did not understand or use mobile

Table 1 Socio demographic characteristics of diabetes mellitus patients in a low-income country, 2022 (n=398)

Variables	Category	Frequency (N)	Percentage (%)
Age (year)	<30	91	22.9
	30–45	155	38.9
	>45	152	38.2
Gender	Male	224	56.3
	Female	174	43.7
Place of residence	Urban	276	69.3
	Rural	122	30.7
Educational status	Unable to read and write	56	14.1
	Informal educational	109	27.4
	Primary school	27	6.8
	Secondary school	66	16.6
	Higher education and above	140	35.2
Marital status	Single	48	12.1
	Married	249	62.6
	Separated	58	14.6
	Others	43	10.7
Occupational status	Housewife	59	14.8
	Government employed	97	24.4
	Non-government employed	38	9.5
	Farmer	45	11.3
	Merchant	103	25.9
	Student	31	7.8
	Others*	25	6.3
Religion	Muslim	102	25.6
	Protestant	103	25.8
	Orthodox	168	42.3
	Catholic	25	6.3
Time to reach the health facility	<1 hour	306	76.9
	≥1 hour	92	23.1
Income status (ETH Birr)	<1500	20	5.0
	1500–3000	103	25.9
	>3000	275	69.1

Others: separated and divorced.
*Daily labourer and unemployed.

applications easily. A total of 247 (67.1%) of the patients used the internet on mobile devices (table 3).

Willingness to use mHealth applications

A total of 284 (71.4%) of respondents were willing to use mobile health applications with a 95% CI (66.8% to 75.9%). A total of 209 (52.5%) respondents had a favourable attitude towards using mobile health applications, and 293 (73.6%) of the participants perceived the usefulness of mobile health applications for self-care management. Moreover, 281 (70.6%) of the participants perceived the mobile health application to manage DM as easy to use (figure 1).

Factors associated with willingness of patients with DM to use the mHealth application

Results of the bivariate analyses showed that age, place of residence, time to reach the facility, use of the internet on mobile devices, substance use, time to reach the health facility, own mobile device, diabetes follow-up time, obtained education during follow-up, obtained medication at any time, attitude, perceived ease of use and perceived usefulness were associated with willingness to use mHealth applications at a p<0.2. All of these associated factors were entered in the multivariable logistic regression analysis model to control for the effect of confounders (table 4).

Table 2 Clinical and behavioural characteristics of patients with diabetes mellitus (DM) in a low-income country, 2022 (n=398)

Variables	Category	Frequency (N)	Percentage (%)
DM types	Type 1	149	37.4
	Type 2	249	62.6
Have comorbidity	Yes	193	48.5
	No	205	51.5
Comorbidity type	Hypertension	53	27.5
	Cardiovascular	71	36.8
	Kidney failure	26	13.5
	Neuropathy	30	15.5
	Others	13	6.7
Route of medication	Pill	236	59.3
	Injection	162	40.7
Time since diagnosis	<1 year	70	17.6
	1–3 years	153	38.4
	>3 years	175	44.0
Diabetes follow-up time	<1 year	88	22.1
	1–3 years	108	27.1
	>3 years	202	50.8
Obtained education during follow-up	Yes	108	27.1
	No	290	72.9
Obtained medication regularly	Yes	302	75.9
	No	96	24.1
Substance use	Yes	85	21.4
	No	313	78.6
Type of substance use	Alcohol	18	21.2
	Khat	43	50.6
	Cigarette	24	28.2
A habit of excessive sugar consumption (per week)	Yes	113	28.4
	No	285	71.6
A habit of physical exercise (per week)	Yes	156	39.2
	No	242	60.8
Frequency of physical exercise (per week)	Once	77	49.3
	Twice	58	37.2
	Three times	12	7.7
	Every day	9	5.8

Others: cancer, retinopathy and stroke.

The multivariable logistic regression model identified the age of patients, place of residence, use of the internet on mobile devices, attitude, perceived ease of use and perceived usefulness as being associated with willingness to use mHealth applications at a p<0.05 (table 4).

As the result summarised in table 4 shows, patients with DM below 30 years of age were 2.21 times more likely to be willing to use mobile health applications (AOR 2.21; 95% CI 1.22 to 4.10) than those who were 35 years of age or older after controlling other variables. Patients who were urban residents were 2.12 times more likely to be willing to use mobile health applications (AOR 2.12; 95% CI 1.12 to 3.98) than those who lived in rural areas, keeping other variables constant. Patients who accessed the internet on mobile devices were 3.91 times more likely to be willing to use mobile health applications (AOR 3.91; 95% CI 1.31 to 11.5) as compared with their counterparts. Similarly, patients who had a favourable attitude towards mobile health applications to manage DM were 5.20 times more likely to be willing to use mobile health applications (AOR 5.20; 95% CI 2.60 to 10.40).

In addition, patients who perceived mHealth applications as easy were 2.57 times more likely to be willing to use mobile health applications (AOR 2.57; 95% CI 1.34 to 4.85) as compared with patients who did not perceive ease of use. Similarly, patients who perceived mHealth

Table 3 Pattern of mobile devices utilisation among patients with diabetes mellitus in a low-income country, 2022 (n=398)

Variables	Category	Frequency (N)	Percentage (%)
Own mobile device	Yes	298	74.9
	No	100	25.1
Type of mobile device	Regular/standard	104	27.5
	Smartphone	216	57.2
	Tablet	27	7.1
	PC	31	8.2
The mobile app used for disease management	Yes	52	13.1
	No	346	86.9
Can you understand how to use the mobile app easily	Yes	127	31.9
	No	271	68.1
Use the internet on mobile devices	Yes	247	62.1
	No	151	37.9

PC, personal computer.

applications as useful were 4.67 times more likely to be willing to use mobile health applications (AOR 4.67; 95% CI 1.95 to 5.77) as compared with patients who did not perceive the usefulness of the mHealth applications (table 4).

DISCUSSION

The purpose of this study was to assess willingness of patients with DM to use mobile health applications and associated factors for self-care management. The result showed that willingness to use mHealth applications among patients with DM in the Oromia region was high (71.4%, 95% CI (66.8% to 75.9%)). This result was in line with the willingness to use mHealth services in Ethiopia.[1] However, This finding was higher than London (63.9%),[20] USA (66.9%),[21] China (66.1%).[22] This difference might be due to sample size, data collection technique (the self-administered method in China), study period and

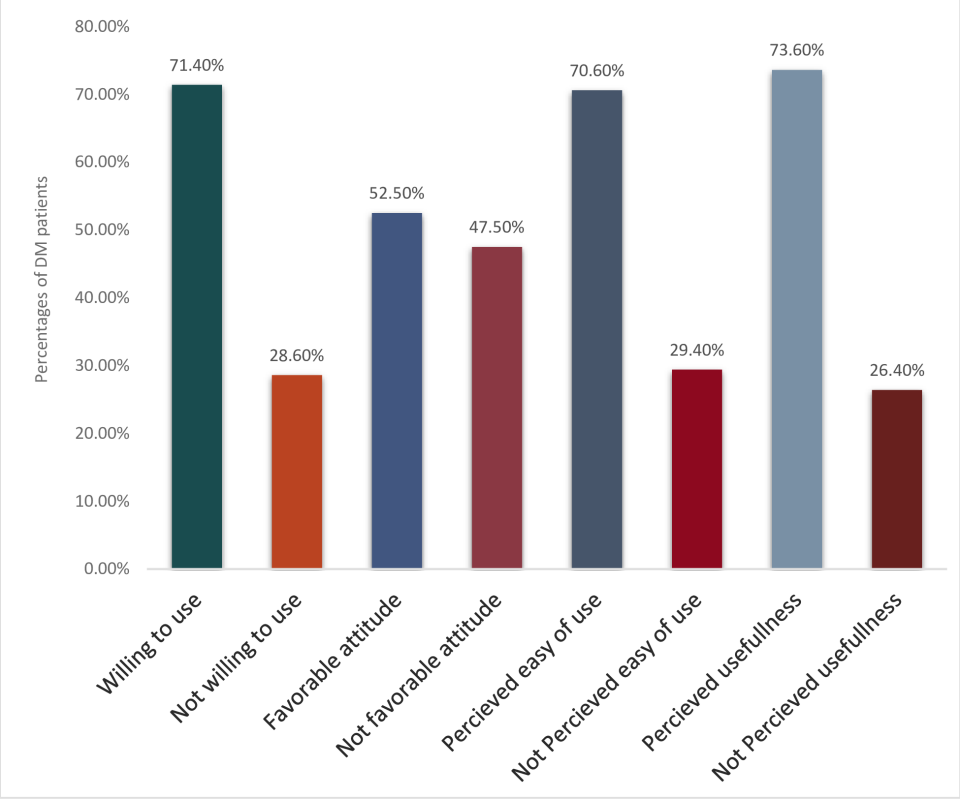

Figure 1 Willingness to use mobile health applications among patients with diabetes mellitus (DM) in southwest public hospitals, Ethiopia, 2022. n=(398).

Table 4 Bivariable and multivariable binary logistic regression analysis of factors associated with willingness of patients with DM to use mHealth application in a low-income country, 2022

Variable	Category	Willingness		Crude OR (95% CI)	Adjusted OR (95% CI)	P value
		Yes	No			
Age (year)	<30	59	32	0.70 (0.04 to 0.94)	2.21 (1.22 to 4.10)	0.03*
	30–45	115	40	1.10 (1.03 to 3.59)	1.51 (0.80 to 2.81)	0.43
	>45	110	42	1	1	
Place of residence	Urban	189	87	0.62 (0.34 to 2.64)	2.12 (1.12 to 3.98)	0.02*
	Rural	95	27	1	1	
Time to reach the facility	<1 hour	226	80	1	1	
	>1 hour	38	54	4.01 (3.21 to 5.85)	0.60 (0.31 to 1.10)	0.61
Own mobile device	Yes	226	72	2.30 (1.61 to 4.22)	1.07 (0.44 to 2.60)	0.32
	No	58	42	1	1	
Internet access	Yes	164	83	0.51 (0.28 to 2.66)	3.91 (1.31, 11.5)	0.03*
	No	120	31	1	1	
Substance use	Yes	45	40	1	1	
	No	239	74	0.35 (0.21, 0.57)	1.63 (0.98 to 2.70)	0.77
Obtained education during follow-up	Yes	60	48	0.37 (0.14 to 5.64)	2.74 (1.25 to 6.04)	0.34
	No	224	66	1	1	
Obtained medication regularly	Yes	204	98	1.92 (0.23 to 2.59)	1.06 (0.49 to 2.26)	0.11
	No	50	46	1	1	
Diabetes follow-up time	<1 year	62	26	1	1	
	1–3 years	63	45	1.71 (1.10 to 2.90)	1.29 (0.71 to 2.40)	0.72
	>3 years	139	63	1.10 (1.59 to 4.21)	2.03 (0.94 to 4.34)	0.21
Attitude	Favourable	166	43	2.32 (2.90 to 5.50)	5.20 (2.60 to 10.40)	0.01*
	Not favourable	118	71	1	1	
Perceived ease of use	Easy	223	58	3.53 (2.26 to 6.25)	2.57 (1.34 to 4.85)	0.002†
	Not easy	61	56	1	1	
Perceived usefulness	Useful not	220	73	1.93 (1.24 to 7.18)	4.67 (1.95 to 5.77)	0.001†
	Useful	64	41	1	1	

*Statistically significant at p<0.05,
†Statistically significant at p<0.01.
CI, Confidence Interval; DM, diabetes mellitus; OR, Odd Ratio.

sociodemographic differences between patients. It is also lower than studies about willingness or intention to use mHealth services among reproductive women done in Ethiopia (78.9%),[16] China (79.5%), USA (80.6%).[23] The discrepancy might be the result of the socioeconomic differences between the participants, or the differences in the ICT development and study period.

Moreover, this study identified numerous factors, including age, place of residence, intermate usage, attitude, perceived usefulness and perceived ease of use that were associated with willingness to use a mHealth app in southwest Ethiopia.

Patients with DM below 30 years of age were associated with a willingness to use mobile health applications. This finding was in line with different studies.[20 22] The possible reason for this is that young person is close to new technologies and accept them easily. Furthermore, a mobile health app that would be crucial for the young age group rather than the older age. As was to be expected, there was high usage of technology and mobile devices, which confirms the potential for and necessity of creating interventions that make use of smartphones as a platform for young people's health. Younger people can explain why they use particular apps, why they manage them, and what qualities and features health apps should have.

Patients who were urban residents were 2.12 times more likely to be willing to use mobile health applications (AOR 2.12; 95% CI 1.12 to 3.98) than those who lived in rural areas. This finding is consistent with different study settings.[24 25] This finding could be explained by the fact that urban residents were more likely than rural residents to have access to mobile health devices, internet access and interactive new technology. Moreover, in rural areas, the challenges range from the high price of mobile

devices, lack of knowledge about their use, a lack of infrastructure or both.

Patients who access the internet on mobile devices were 3.91 times more likely to be willing to use mobile health applications (AOR 3.91; 95% CI 1.31 to 11.5) as compared with their counterparts. This finding is consistent with different study settings.[22 25] This could be due to patients' increased use of the internet on their mobile devices, improved access to health information and increased awareness of the use of mobile health applications for self-care management. Therefore, expanding internet connectivity and infrastructure is necessary for the effective distribution of mHealth applications for diabetes patients' self-care management.

Patients who had a favourable attitude towards mobile health applications to manage DM were 5.20 times more likely to be willing to use mobile health applications (AOR 5.20; 95% CI 2.60 to 10.40). This finding is consistent with the findings of other studies.[26 27] This demonstrates how patients' attitudes towards mobile health technologies influence their willingness to use them positively. Patients' intentions to use mobile health applications also increase as they view these technologies as a tool to improve their health management. The probable cause of this is that patients with DM who have a settled, positive attitude towards mobile health applications will be extremely incensed by new solutions. Therefore, a strong emphasis should be placed on activities that improve attitudes, such as computer access, ongoing training and support, and knowledge sharing on mobile health technology for self-care management.

Patients who perceived mHealth applications as easy were 2.57 times more likely to be willing to use mobile health applications (AOR 2.57; 95% CI 1.34 to 4.85), as compared with patients who did not perceive ease of use.[16 28 29] Users tend to focus more on a system's usability when they have no or little prior experience with it. This suggests that if a new mHealth service is perceived as being difficult to use, users will not use it, regardless of how valuable the system may be. Users will stop using mHealth solutions that are not user-friendly, according to research. If a person believes that a new system will benefit them, any difficulty with using it may be resolved. Deploying mHealth applications may, therefore, need further education on how to use and manage the new system for better adoption.[16]

Patients who perceived mHealth applications as useful were 4.67 times more likely to be willing to use mobile health applications (AOR 4.67; 95% CI 1.95 to 5.77) as compared with patients who did not perceive the usefulness of the mHealth applications.[28 29] This showed that no matter how simple or complex the mobile health application is to use, patients will not use it if they do not believe it will be valuable to them. As a result, it is important to make sure that the system will enhance the desired health results when developing it.

Limitations of the study and future research

The study has some limitations, the study was interviewer based, so the responses might have been affected by bias introduced by the interviewers, the study did not include private hospitals; and the study used quantitative approaches, which may affect the generalisability of the findings. Hence, future research studies could include patients with diabetes from private hospitals, support the finding with a qualitative study and use a health information technology acceptance model, so that the results could be generalisable.

CONCLUSION

Overall, this study showed that the proportion of willingness of patients with DM to use mobile health applications for self-care management was high. The willingness of patients with DM to use mobile health applications was influenced by their age, place of residence, internet access, attitude, perceived ease of use and perceived usefulness. Considering these factors could provide insight for designing and implementing diabetes management applications on mobile devices in Ethiopia.

Author affiliations
[1]Department of Health Informatics, College of Health Science, Mattu University, Mettu, Ethiopia
[2]Department of Epidemiology and Biostatistics, Institute of Public Health, College of Medicine and Health Science, University of Gondar, Gondar, Ethiopia
[3]Department of Health Informatics, Institute of Public Health, College of Medicine and Health Science, University of Gondar, Gondar, Ethiopia
[4]Department of Health System Management and Health Economics, School of Public Health, College of Medicine and Health Science, Bahir Dar University, Bahir Dar, Ethiopia

Acknowledgements The authors would like to thank Mettu University's ethical committee for the approval of ethical clearance. Also, the authors are grateful to the Oromia region public hospital management, data collectors, supervisors and study participants.

Contributors ADW: conception and design, study selection, data curation, analysis and interpretation of data, funding acquisition, investigation, methodology. AAS and AWD: supervision, validation, visualisation and reviewing.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Consent obtained directly from patient(s).

Ethics approval This study involves human participants and was the ethical clearance was approved by Mettu University's ethical review committee with approval number MeU/RAC/1036/15. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

ORCID iDs
Agmasie Damtew Walle http://orcid.org/0000-0001-9583-5876
Adamu Ambachew Shibabaw http://orcid.org/0000-0002-0207-4746

Sisay Maru Wubante http://orcid.org/0000-0001-9442-5074
Chalachew Msganaw Yehula http://orcid.org/0000-0003-2989-8128
Addisalem Workie Demsash http://orcid.org/0000-0002-9356-8126

REFERENCES

1 Jemere AT, Yeneneh YE, Tilahun B, *et al*. Access to mobile phone and willingness to receive mHealth services among patients with diabetes in Northwest Ethiopia: a cross-sectional study. BMJ Open 2019;9:e021766.
2 Federation I. IDF diabetes atlas, tenth. International Diabetes 2021.
3 Dereje N, Earsido A, Temam L, *et al*. Prevalence and associated factors of diabetes mellitus in Hosanna town, southern Ethiopia. Annals of Global Health 2020;86.
4 Sun H, Saeedi P, Karuranga S, *et al*. IDF diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 2022;183:S0168-8227(21)00478-2:109119.:.
5 Shibuta T, Waki K, Tomizawa N, *et al*. Willingness of patients with diabetes to use an ICT-based self-management tool: a cross-sectional study. BMJ Open Diab Res Care 2017;5:e000322.
6 Olamoyegun MA, Raimi TH, Ala OA, *et al*. Mobile phone ownership and willingness to receive mHealth services among patients with diabetes mellitus in South-West, Nigeria. *Pan Afr Med J* 2020;37:29.
7 Varshney U. Mobile health: four emerging themes of research. Decis Support Syst 2014;66:20–35.
8 Palos-Sanchez PR, Saura JR, Rios Martin Miguel Ángel, Rios Martin MÁ, *et al*. Toward a better understanding of the intention to use mHealth Apps: exploratory study. *JMIR Mhealth Uhealth* 2021;9:e27021.
9 Geirhos A, Stephan M, Wehrle M, *et al*. Standardized evaluation of the quality and persuasiveness of mobile health applications for diabetes management. Sci Rep 2022;12:1–10.
10 Silver L. Smartphone ownership is growing rapidly around the world, but not always equally 2019.
11 Wang Y, Wu T, Chen Z. Active usage of mobile health applications: cross-sectional study. *J Med Internet Res* 2021;23:e2533012.
12 Shan R, Sarkar S, Martin SS. Digital health technology and mobile devices for the management of diabetes mellitus: state of the art. Diabetologia 2019;62:877–87.
13 Waki K, Fujita H, Uchimura Y, *et al*. DialBetics: a novel smartphone-based self-management support system for type 2 diabetes patients. J Diabetes Sci Technol 2014;8:209–15.
14 Gücin NÖ, Berk ÖS. Technology acceptance in health care: an integrative review of predictive factors and intervention programs. *Procedia - Social and Behavioral Sciences* 2015;195:1698–704.
15 Lim S, Kim S-Y, Kim JI, *et al*. A survey on ubiquitous healthcare service demand among diabetic patients. *Diabetes Metab J* 2011;35:50.
16 Mekonnen ZA, Gelaye KA, Were MC, *et al*. Mothers intention and preference to use mobile phone text message reminders for child vaccination in Northwest Ethiopia. *BMJ Health Care Inform* 2021;28:e100193.
17 Bogale B, Habte A, Haile D, *et al*. Willingness to receive mHealth messages among diabetic patients at Mizan Tepi university teaching hospital: implications for digital health. Patient preference and adherence 2022;16:1499–1509.
18 Gagnon MP, Orruño E, Asua J, *et al*. Using a modified technology acceptance model to evaluate healthcare professionals' adoption of a new telemonitoring system. Telemed J E Health 2012;18:54–9.
19 Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13:319–40.
20 Goedel WC, Mitchell JW, Krebs P, *et al*. Willingness to use mobile phone apps for HIV prevention among men who have sex with men in London: web-based survey. JMIR mHealth and uHealth 2017;5:e153.
21 Hargittai E, Redmiles EM, Vitak J. Americans' willingness to adopt a COVID-19 tracking App. *FM* 2020;25:11.
22 Yang S, Chen Y, Zhou L, *et al*. Willingness to adopt Mhealth among Chinese parents during the COVID-19 outbreak: Cross-sectional questionnaire study. *JMIR Mhealth Uhealth* 2021;9:e23155.
23 Abelson JS, Symer M, Peters A, *et al*. Mobile health apps and recovery after surgery: what are patients willing to do? *Am J* Surg 2017;214:616–22.
24 Basu S, Garg S, Kumar R, *et al*. The willingness for using mobile phone for health education among women caregivers of under 5 children in an urban resettlement colony in Delhi, India. *Indian J Community Health* 2017;29:439–44.
25 Reddy MM, Thekkur P, Majella MG, Selvaraj K, *et al*. Use of mobile phone in healthcare: readiness among urban population of Puducherry, India. Int J Med Public Health 2016;6:94–7.
26 Huang C-Y, Yang M-C. Empirical investigation of factors influencing consumer intention to use an artificial Intelligence-Powered mobile application for weight loss and health management. Telemedicine and e-Health 2020;26:1240–51.
27 Tobias G, Sgan-Cohen H, Spanier AB, *et al*. Perceptions and attitudes toward the use of a mobile health APP for remote monitoring of gingivitis and willingness to pay for mobile health Apps (Part 3): mixed methods study. JMIR Form Res 2021;5:e26125.
28 Chen Y, Xu Q. The willingness to use mobile health and its influencing factors among elderly patients with chronic heart failure in Shanghai, China. *International Journal of Medical Informatics* 2022;158:104656.
29 Deng Z, Hong Z, Ren C, *et al*. What predicts patients' adoption intention toward mHealth services in China: empirical study. JMIR Mhealth Uhealth 2018;6:e172.

BMJ Health & Care Informatics

Prevalence of electronic screening for sepsis in National Health Service acute hospitals in England

Kate Honeyford ,[1] Amen-Patrick Nwosu ,[2] Runa Lazzarino,[3] Anne Kinderlerer,[4] John Welch,[5] Andrew J Brent,[6,7] Graham Cooke,[4,8,9] Peter Ghazal,[10] Shashank Patil,[11] DiAIS Co-investigators, Ceire E Costelloe[1,12]

To cite: Honeyford K, Nwosu A-P, Lazzarino R, *et al*. Prevalence of electronic screening for sepsis in National Health Service acute hospitals in England. *BMJ Health Care Inform* 2023;**30**:e100743. doi:10.1136/bmjhci-2023-100743

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100743).

Received 30 January 2023
Accepted 12 April 2023

For numbered affiliations see end of article.

Correspondence to
Dr Kate Honeyford;
kate.honeyford@icr.ac.uk

ABSTRACT

Sepsis is a worldwide public health problem. Rapid identification is associated with improved patient outcomes—if followed by timely appropriate treatment.

Objectives Describe digital sepsis alerts (DSAs) in use in English National Health Service (NHS) acute hospitals.

Methods A Freedom of Information request surveyed acute NHS Trusts on their adoption of electronic patient records (EPRs) and DSAs.

Results Of the 99 Trusts that responded, 84 had an EPR. Over 20 different EPR system providers were identified as operational in England. The most common providers were Cerner (21%). System C, Dedalus and Allscripts Sunrise were also relatively common (13%, 10% and 7%, respectively). 70% of NHS Trusts with an EPR responded that they had a DSA; most of these use the National Early Warning Score (NEWS2). There was evidence that the EPR provider was related to the DSA algorithm. We found no evidence that Trusts were using EPRs to introduce data driven algorithms or DSAs able to include, for example, pre-existing conditions that may be known to increase risk.

Not all Trusts were willing or able to provide details of their EPR or the underlying algorithm.

Discussion The majority of NHS Trusts use an EPR of some kind; many use a NEWS2-based DSA in keeping with national guidelines.

Conclusion Many English NHS Trusts use DSAs; even those using similar triggers vary and many recreate paper systems. Despite the proliferation of machine learning algorithms being developed to support early detection of sepsis, there is little evidence that these are being used to improve personalised sepsis detection.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Digital alerts are being introduced into hospital systems in England as they switch to electronic patient records (EPRs). Little is known about the presence of digital sepsis alerts in these hospitals or the accuracy of the underlying algorithms.

WHAT THIS STUDY ADDS

⇒ The majority of hospitals with EPRs use digital sepsis alerts, with National Early Warning Score 2 being the most common algorithm to detect all-cause deterioration including sepsis. The algorithm in use is influenced by the EPR contracted by the Trust.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Detailed patient data within EPRs is not currently exploited to improve digital sepsis alerts in hospitals. We recommend that NHS organisations are open about the digital tools in use and their effectiveness rigorously evaluated.

INTRODUCTION

Sepsis is a worldwide public health problem, with a recent report estimating a 11 million global death toll in 1 year alone. Early diagnosis and management is crucial to improve patient outcomes,[1 2] with inconsistent recognition and management of sepsis being repeatedly highlighted as a safety concern in hospital service/quality of care audits.[3]

These related issues currently make early sepsis recognition more challenging: interindividual heterogeneity in the underlying aetiology and clinical phenotype; inconsistency in the implementation of a consensus clinical definition; and most critically, the lack of a reliable test for sepsis.[4]

Screening for sepsis is widely implemented across countries, and is essential for prompt treatment and optimal outcomes.[5] Latest international guidelines recommend that all hospitals and healthcare systems adopt sepsis performance improvement programmes, which include the use of screening tools to promptly identify sepsis.[1 6] However, compliance with these guidelines is not universal, and implementation is an ongoing challenge.[7]

Currently, hospitals in England are required to screen both emergency department (ED) patients and inpatients for sepsis 'where appropriate' and there have been associated financial incentives towards this.[8] Recent guidelines are summarised in figure 1. To date, none of these guidelines considers the

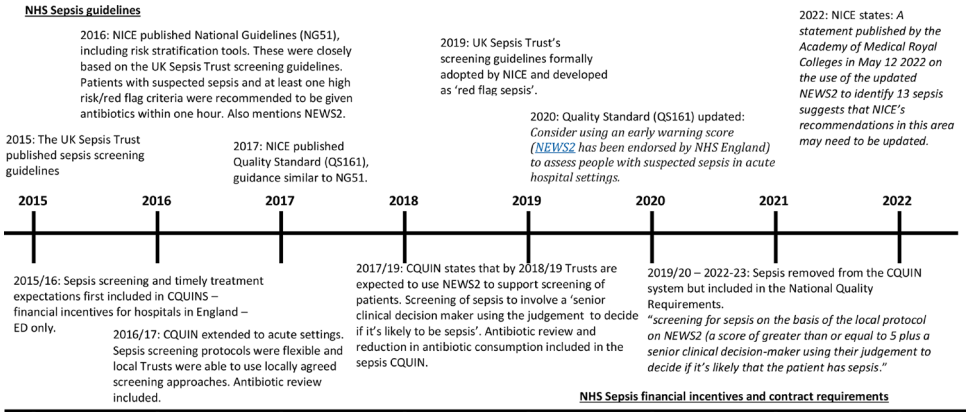

Figure 1 Timeline showing the development of sepsis guidelines and incentives in the NHS in England. NEWS2, National Early Warning Score version details of NICE guidelines and quality Standards are available at: www.nice.org.uk/guidance/conditions-and-diseases/infections/sepsis, details of CQUINs are available at: www.england.nhs.uk/nhs-standard-contract/cquin/, details of NHS standard contract are available at: www.england.nhs.uk/nhs-standard-contract/previous-nhs-standard-contracts/. CQC, Care Quality Commission; CQUIN, Commissioning and Quality Innovation; ED, emergency department.; NICE, National Institute of Clinical Excellence; NHS, National Health Service

use of electronic tools to aid screening, or their potential advantages and disadvantages.

Despite their absence from current guidelines, electronic screening tools for sepsis have been in use in English hospitals for over 5 years. Previous work from our group showed that the introduction of a digital sepsis screening tool and accompanying alert was associated with reduction in risk of mortality, and an increase in timely treatment with antibiotics.[9] Individual Trusts have identified improvements in patient outcomes including reductions in septic shock in under 45s from 60% to 7.7%,[10] 70% increase in patients diagnosed with sepsis receiving antibiotics within the target time frame, and 64 potential lives saved 1 year.[11] However, these claims have not been peer reviewed, or adjusted for underlying trends and casemix.

Currently, most electronic screening tools for sepsis available in England are rule based, track and trigger (T&T) systems, that is, systems which rely on periodic observation of selected physiological signs with predetermined criteria for escalating care.[12] The most commonly available tools include systemic inflammatory response syndrome (SIRS) criteria, quick Sepsis-related Organ Failure Assessment (qSOFA), modified Early Warning Scores and, in the UK, National Early Warning Score (NEWS)2.[6] SIRS and qSOFA were initially developed as diagnostic tools for sepsis, but are now commonly used for highlighting patients at risk of poor outcomes from sepsis (details are shown in table 1).[4] These tools often have high sensitivity, but low specificity.[6] The criteria of these tools are applicable to adults and are not directly appropriate for neonates, children or maternity patients; consequently, this paper focuses on digital sepsis tools for use in adults.

Current UK adult sepsis guidelines recommend using NEWS2 (see figure 1) to identify patients at risk of deterioration and then involve a senior clinical decision maker to determine if sepsis is driving the deterioration. This is a simple approach that can easily be linked to electronic systems. None of these algorithms, including NEWS2, makes use of the granular nature of electronic patient records (EPRs); for example, pre-existing conditions and treatments or deviations in vital signs from the normal for an individual patient. This is despite published studies highlighting the benefits and high predictive performance of algorithms based on machine learning approaches which can factor in more detailed patient information.[13] These studies were not conducted in hospital settings, hence evidence of positive results in hospital settings is still limited.[14] Indeed, few digital sepsis alerts (DSAs) available to hospitals have been evaluated in terms of patient benefit as opposed to predictive accuracy.

As the UK National Health Service (NHS) seeks to become paperless and embraces digital technology, the incorporation of digital alerts embedded within the EPR is an attractive option to aid clinical decision-making, and has the potential to increase the quality, efficiency and cost-effectiveness of sepsis care. However, little is known about the digital alerts currently in use or the rationale for their inclusion in healthcare systems. In the case of sepsis, there is some emerging evidence of the effectiveness of these tools, but there are no validated digital tools available to NHS Trusts which have been shown to be effective in improving patient outcomes in a range of settings, nor has there been a recent comprehensive review of the algorithms in use.

In this paper, we describe DSAs, based on English NHS Trusts responses to Freedom of Information (FOI) request.

METHODS

Working with a group of close collaborators identified to be part of a wider project (see box 1 for further details), we identified key aspects of algorithms in use in five NHS hospitals to inform our further work. We used an FoI

Table 1 Summary of key criteria used in sepsis detection algorithms

Clinical observation	qSOFA*	NEWS2 Consider infection			Red Flag NEWS3+possible infection One from lists below		SIRS*	St John sepsis algorithm
		1	2	3	Red	Amber		
Respiration Rate	³22		21–24	≥25	25	21–24	>20	>21
		9–11		<8				
O_2 Saturation	–	94–95	92–93	<91	<91	–	–	–
Supplemental O_2	–		Yes		>92 with supp O_2	–	–	–
Temperature, °C	–	35.1–36.1		<35	–	<36	<36.0 or >38.0	<36.0 or >38.3
		38.1–39.0	³39.1					
Systolic blood pressure	<100	101–110	91–100	<90	<90	91–100		<90
				>220				
Heart rate	–	41–50		<40	>130	91–130	>90	>95
		91–110	111–130	³131				
Level of consciousness	GCS<13 or AMS			ACVPU	ACVPU	Concern	–	–
Lactate	–	–	–	–	³2.0 mmol	–	–	³2.0 mmol
Glucose mg/dL	–	–	–	–	–	–	–	>140 and <200
Fluid balance	–	–	–	–	No urine 18+ hours or <0.5 mL/kg/hour	No urine 12–18 hours	–	–
White cell count	–	–	–	–	–	–	>12 k or <4 k or band >10%	>12 k or <4 k or band >10%
Creatinine mg/dL (µmol/L)	–	–	–	–	–	–	–	Increase of 0.5 over 72 hours
Bilirubin mg/dL	–	–	–	–	–	–	–	<2.0 and < 10.0

*SIRS and qSOFA have previously been included in sepsis diagnostic criteria; in this report, we focus on the use of these tools as prediction/alerting algorithms.
ACVPU, alert, confusion, voice, pain, unresponsive; AMS, altered mental status; GCS, Glasgow Coma Score; NEWS, National Early Warning Score; qSOFA, quick Sepsis-related Organ Failure Assessment ; SIRS, systemic inflammatory response syndrome.

Box 1 DiAIS—Digital Alerting for Sepsis

The DiAIS study is investigating the impact of DSAs on patient outcomes and staff activity in six NHS hospital Trusts across England and Wales.
The implementation of digital alerts in hospitals is a complex health intervention. Therefore, we are using a mixed-methods approach to ensure understanding of the relationship between inherent aspects of the alerts, such as the underlying algorithm and the method of clinician notification. Using appropriate qualitative and quantitative methods, based on the analysis of natural experiments, we will evaluate the implementation of alerts across six NHS Trusts, most of which have adopted distinctive digital alerts.
Outcomes will include in-hospital mortality within 30 days, transfers to the intensive care unit, length of stay and administration of intravenous antibiotics. We will also consider unintended consequences related to unnecessary and inappropriate use of antibiotics.
DiAIS is funded by National Institute for Health and Care Research (NIHR) Health and Social Care Delivery Research (HS&DR) and is working in collaboration with NIHR-Health Informatics Collaborative.

DSAs, digital sepsis alerts.

request to survey all hospitals and used internet searching to gather additional information.

The FoI request was submitted to all acute NHS hospital Trusts in England that have an ED (with the exception of the five NHS hospitals in the pilot work), to collect information on EPRs; electronic sepsis screening tools and the underlying algorithms they use; the association between the underlying algorithm and the alerts to clinicians; the timing of introduction of the electronic screening tool in the hospital; and which staff groups see and respond to alerts. We did not give a definition of EPR in our request which enabled trusts to respond how they deemed most appropriate. The FoI request is available in online supplemental appendix 1.

The results of the FoI were screened by two authors (KH and A-PN). Where there was ambiguity in the response by the trust, for example, if the response indicated that there was no DSA but details of the algorithm and process were supplied, we discussed the response and reached a consensus approach.

RESULTS

FOI of requests were sent to 120 Acute NHS Trusts which had EDs. Responses were received from 94 NHS Trusts. Additional information was gathered from the Digital Alerts for Sepsis (DiAlS) clinical team and from five NHS Trusts participating in DiAlS (see box 1). Of the 99 Trusts for which information was available, 14 (14%) responded that they did not have an electronic health record or EPR. Eighty-four (85%) Trusts responded that they had an EPR. The most common single provider was Cerner (18 Trusts, 21%). System C, Dedalus and Allscripts Sunrise were also relatively common (13%, 7% and 10%, respectively). Four Trusts used Epic and two used in-house systems. Over one-fifth of Trusts (22%) identified a mix of companies providing their EPRs, with EDs and inpatient wards sometimes using different systems, and some identifying various patient administration systems. Further details are provided in table 2. One Trust refused to provide information on their provider citing potential cyberattacks as justification.

DSAs were reported to be in use in 59 of the 85 digital Trusts (69%). Systems based on NEWS2 were the most used across all systems (46 Trusts (78%)). Of these, 29 used a combined approach, an aggregate score of 5 or above or a single parameter of 3 or above, compared with 21 which either specified that they use a score of 5 or above or did not specify. Within Trusts which use NEWS2,

Table 2 EPR providers, digital sepsis alerts and associated algorithms

		NEWS2								
	All EPRs	Sepsis alerts	Alone	& qSOFA	& Red Flag	& sepsis screen	& SIRS	qSOFA alone	Red Flag alone	SIRS alone
Total	**84**	**59**	**22**	**9**	**6**	**8**	**1**	**3**	**8**	**2**
Mix	19 (23%)	14 (24%)	4	3	2	3			2	
Cerner	18 (21%)	15 (27%)	6		2	–	1		4	2
System C	11 (13%)	8 (14%)	1	5		–		1		
Allscripts Sunrise	6 (7%)	3 (5%)	1			1		1	1	
Dedalus	8 (10%)	3 (5%)	1		1	1				
In-house	2 (5%)	1 (2%)	–			1				
Epic	4 (5%)	3 (5%)	2			1				
Other	14 (17%)	8 (15%)	5	1	1	–		1	1	
Missing	2 (2%)	–	1			1				

Percentages are column percentages, showing the proportion of each algorithm associated with each of the main EPR provider.
EPR, electronic patient record; qSOFA, quick Sepsis-related Organ Failure Assessment; SIRS, systemic inflammatory response syndrome.

24 (52%) use an additional screening tool; these include Red Flag Sepsis[15] and qSOFA.[4] A further eight NEWS2 Trusts use an additional screening tool, such as asking for an indication of infection, and one uses SIRS criteria. Some Trusts indicated that their digital system prompts a question or a series of questions about the possibility of sepsis, but there was no evidence that the responses to these were precompleted by the electronic system, despite some of this information being available within the EPR. Eight Trusts responded that they used Red Flag criteria as a stand-alone assessment, while three used qSOFA and two SIRS.

Five Trusts did not give sufficient information to determine the algorithm behind the sepsis alert. An additional four were unwilling to provide information on the algorithm, and three trusts use bespoke systems which were a modification of NEWS2 or Red Flag Sepsis.

We saw some patterns in the algorithm used and the EPR provider. These are summarised in table 2. The EPR provider for all Trusts which use a combination of NEWS2 and qSOFA is system C and for those that use an SIRS-based system it is Cerner. Cerner was also a common provider for Trusts using Red Flag Sepsis alone or in combination with NEWS2.

Willingness to disclose information

As identified above, not all NHS Trusts were happy to disclose information and some Trusts were unable to provide information. Some of these responses are provided in box 2.

Box 2 Example of responses from trusts which were unable or unwilling to provide information on EPR provision or digital sepsis alert algorithms.

'The Trust considers this question to be exempt from disclosure in accordance with section 43.2 of the Freedom of Information Act as to release this information would, or would be likely to, prejudice the commercial interests of the supplier.'
'Care Flow Vitals Clinical uses qSOFA scoring for detection of patients at risk of sepsis. The exact algorithm is not known by the Trust.'
'N/K'.
'"Unable to provide as this is managed by the supplier'.
'In view of cybersecurity attacks on organisations, the Trust considers that public release of this information could put the Trust's system and information contained on that system at risk. Accordingly the Trust considers, therefore, that section 31(1) of the Freedom of Information Act 2000 applies. As this is a qualified exemption, the Trust has applied the public interest test as required and deems that, on balance and for the reason stated above, the public interest lies in not disclosing this information.'
'Unable to provide as the algorithms are part of a third party system and proprietary knowledge'.

EPR, electronic patient record.

DISCUSSION

The majority of Trusts responding reported having an EPR and over 20 different providers were identified as operational in NHS Trusts in England. Three-quarters of digital Trusts responded that they had a DSA and most these use NEWS2 as part of their sepsis alerting system. This is the approach included in the NHS National Standard Contract.[16]

There is evidence that the provider of the EPR in use in the hospital is associated with the underlying sepsis algorithm in use in the Trust. For example, SIRS-based alerts are only found in Trusts where the provider is Cerner and qSOFA is part of the System C sepsis alert system.

Given NEWS2 is the nationally recommended system for identifying patients who need to be screened for sepsis it is not surprising that NEWS2 is the most commonly used system, usually with a threshold of 5 as the trigger for review and consideration of sepsis. In addition, many Trusts include a score of 3 in any single parameter, despite national guidance moving away from this approach as a trigger for significant escalation as it is a poor predictor of risk.[17] This may be a legacy of the overlap between indicators in Red Flag Sepsis and NEWS2.

There was no evidence that Trusts' digitisation of patient health records was associated with the introduction of more complex algorithms, either data-driven, machine learning-based algorithms or algorithms which were able to include pre-existing conditions or patient information.

Our review of the EPR systems in use in English Trusts are in line with those found by Warren *et al*.[18] In their study of NHS Trusts (2017–2018), 23% of Trusts reported having no electronic system, suggesting an increase in adoption of electronic systems since 2018. Cerner was the most commonly reported provider (18%), then DXC (13%) and System C (11%). DXC were the providers of the Lorenzo EPR in 2017, but were bought out by Dedalus, an Italian-based provider, in April 2021,[19] which was still a common provider in our survey (10%). We found a higher proportion with a mixed system than Warren *et al*, which may reflect a less precise definition of EPR in our study or changes over time.

In 2007, the NIHR reported that 'T&T systems were in widespread use in NHS acute hospitals'; it is therefore no surprise that we have found that the majority of digital trusts are using digital T&T systems as key components of their DSAs.[20] In addition, Trusts are expected to use NEWS2 as a screening system for deteriorating patients. Advantages of T&T systems include the ability to monitor large numbers of patients without 'a large increase in workload', and digital enhancement of these systems has clear advantages, for example, one study showed errors in pen-and-paper T&T systems, with errors in 29% calculated scores reviewed (n=84), half of these led to incorrect clinical action.[12] Although NEWS2 is a relatively simple system, there is an obvious advantage in clinical data being aggregated automatically and removing the need for busy clinical staff determining and totalling the 'points' value of each observation.

In addition to NEWS2, qSOFA and SIRS were relatively commonly used, with adoption associated with the EPR provider. qSOFA was also used in addition to the

expected NEWS2. It is possible to trigger a qSOFA score of 2 or more without aggregating to a score of 5 or more in NEWS2, but there is no evidence that using the two combined leads to improved specificity.[21]

SIRS-based systems were only in use in Trusts where Cerner was the EPR provider. SIRS is not now currently considered a useful way of defining sepsis, but may be useful in predicting poor outcomes for patients.[9 22] SIRS-based algorithms do make use of more detailed information contained within patient records, for example, recent lactate and bilirubin levels which reflect organ dysfunction. In addition, the Cerner-based algorithm can be set different thresholds for patients with diabetes or undergoing dialysis, the only system which automatically considers wider information about the patient. However, most studies suggest that SIRS has limited utility in accurately identifying sepsis[21] or predicting mortality in patients.[23]

NEWS2 and qSOFA were designed to be easily performed at the bedside and qSOFA was developed using a parsimonious model to achieve a 'simple scoring system with the fewest number of variables associated with the greatest predictive ability'.[24] Although this approach makes sense in low-resource settings without EPRs, it does not take advantage of the available granular patient data. This includes information from the current visit, previous contacts with the hospital and potentially information from recent primary care appointments.

In this paper, we have not examined the potential benefits or harms of DSAs; a systematic review[25] did not find a reduction in mortality, in contrast to Honeyford *et al*.[9] Studies have shown improvements in achievement of process measures. In different parts of the hospital, alerts' potential to improve patient outcomes varies; in modern EDs, there is often continuous electronic monitoring. In highly resourced EDs, unwell patients are usually reviewed early by a senior clinician, hence there may be limited value of an alert system. Where staff are under increased pressure, alerts may be more important.[25] This contrasting evidence emphasises the need for robust evidence to determine the most appropriate DSA.

A minority of Trusts were unwilling or unable to give detailed information about their EPR or the underlying algorithm used for their sepsis alert. This is an important aspect of the introduction of digital alerts in hospitals in England/UK. Currently, there is no clear approval process for digital alerts, and, hence, no necessity for hospitals to use 'approved' digital alerts. Two high-profile alerts have recently been identified at best as having no utility, at worst, causing patient harm.[14 26]

Initially, it was hypothesised that Trusts who responded that they had an EPR would be paperless or heavily paper reduced. However, responses to the question indicate some Trusts are combining electronic and paper systems; we were, therefore, unable to determine how many Trusts are paper reduced/less. We had similar challenges in determining the level of 'digital' in the sepsis alert. While some Trusts answered 'yes' to DSAs, examination of the details provided suggested that the alert relied on paper. The UK Sepsis Trusts describes sepsis screening as a two-part process, recommending that patients are 'screened for sepsis' if they have a NEWS2 score of 5 or more. It was difficult to determine whether the DSA described by respondents was the 'prescreen' to identify which patients needed screening for sepsis. Trusts which did not explicitly state that they used NEWS2 are highly likely to be using NEWS2 as part of their sepsis screening system, however, this may not be digital or not be considered part of the DSA. The combined paper and digital model requiring significant staff input to determine the requirement for review reduces some of the advantages of automation.

We opted to use an FOI request to increase response rate. People completing FOIs in NHS Trusts will not necessarily have the knowledge to answer the questions and err on the side of caution. Although some Trusts responded to the FOI with 'not known' or equivalent we are sure that there are staff in the Trust who know the algorithm being used. Finally, a minority of Trusts were unwilling or unable to give detailed information about their EPR or the underlying algorithm used for their sepsis alert. This is an important aspect of the introduction of digital alerts in hospitals in England. Currently, there is no necessity for hospitals to use approved digital alerts and no clear approval process. Two high-profile alerts have recently been identified as having no utility and at worst, causing patient harm.[14 26]

Wong *et al*[14] have highlighted that in the USA 'the ease of integration within the EPR and loose federal regulations' means that hospitals adopt algorithms with ease, without a detailed knowledge of real-world performance. This is also the case in England, however, the Medical Health Regulatory Authority are now recommending that software as a medical device should undergo proper scrutiny, 'commensurate with risk'. There is a need for a strong methodological library for evaluating digital tools, including determining risk. This is the focus of the UK NIHR DiAlS study that is evaluating electronic sepsis screening tools which are currently in use in England.

CONCLUSION

Digital tools currently in use in acute hospitals in England use simple algorithms, based on paper-based T&T systems and are not taking advantage of granular data available in the EPR. While the majority of NHS Trusts in England are using NEWS2, as required in the National Standard Contract, this was not designed as a digital tool nor developed within data rich environments. Many Trusts are using alternative algorithms, often in combination with NEWS2, which do not have a strong evidence base. Studies which compare these approaches are vital to inform on the most effective practice.

As EPRs become universal, there is enormous potential in harnessing granular data to improve the performance of digital tools to support care of deteriorating patients. However, we need a strong methodological

evaluation approach and clinicians and hospital leaders have a responsibility to understand the digital tools in use in their hospitals. We would go further and suggest that there should be a publicly accessible registry of digital alerting tools in use in hospitals, including DSAs.

Author affiliations
[1]Team Health Informatics, Institute of Cancer Research, London, UK
[2]MRC Centre for Global Infectious Disease Analysis, Jameel Institute, School of Public Health, Imperial College London, London, UK
[3]Nuffield Department of Primary Care and Health Sciences, University of Oxford, Oxford, UK
[4]Imperial College Healthcare NHS Trust, London, UK
[5]Critical Care Department, University College Hospital, London, UK
[6]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[7]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[8]Department of Infectious Disease, Imperial College, London, UK
[9]National Institute for Health Research Imperial Biomedical Research Centre, London, UK
[10]Systems Immunity Research Institute, School of Medicine, Cardiff University, Cardiff, UK
[11]Emergency Department, Chelsea and Westminster Healthcare NHS Trust, London, UK
[12]Health Informatics Team, Royal Marsden NHS Foundation Trust, London, UK

Twitter Anne Kinderlerer @akinderlerer and Shashank Patil @DrShashankPatil

Contributors KH, CEC and the DiAIS co-investigators conceived this study. KH and A-PN collated and summarised the data. All author reviewed and edited early drafts of the manuscript. All authors contributed to the article and approved the submitted version. CEC is the guarantor for this study.

Funding KH and RL are supported by the NIHR [HS&DR] Project: NIHR129082. CEC holds an NIHR award (NIHR 129082) which supports this work. GC is supported in part by the NIHR Biomedical Research Centre of Imperial College NHS Trust. PG is funded by Ser Cymru Programme sponsored by Welsh Government and EU-ERDF. A-PN acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. A-PN acknowledges funding by Community Jameel.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Please contact the DiAIS Team for the information in a spreadsheet format. All data are publicly available on individual NHS Trust's websites.

ORCID iDs
Kate Honeyford http://orcid.org/0000-0002-5533-7693
Amen-Patrick Nwosu http://orcid.org/0000-0001-6991-7798

REFERENCES

1 Evans L, Rhodes A, Alhazzani W, *et al*. Executive summary: surviving sepsis campaign: international guidelines for the management of sepsis and septic shock 2021. *Crit Care Med* 2021;49:1974–82.
2 Rudd KE, Johnson SC, Agesa KM, *et al*. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* 2020;395:200–11.
3 Commission CQ. The state of care in NHS acute hospitals: 2014-2016. Care Quality Commission [Internet]. 2017. Available: http://www.cqc.org.uk/sites/default/files/20170302b_stateofhospitals_web.pdf
4 Singer M, Deutschman CS, Seymour CW, *et al*. The third International consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* 2016;315:801.
5 Alberto L, Marshall AP, Walker R, *et al*. Screening for sepsis in general hospitalized patients: a systematic review. *J Hosp Infect* 2017;96:305–15.
6 Schinkel M, Nanayakkara PWB, Wiersinga WJ. Sepsis performance improvement programs: from evidence toward clinical implementation. *Crit Care* 2022;26:77.
7 Vincent J-L, Singer M, Einav S, *et al*. Equilibrating SSC guidelines with individualized care. *Crit Care* 2021;25:397.
8 NHS England Contracting and Incentives Team. *Commissioning for Quality and Innovation (CQUIN) Guidance for 2015/16*. Leeds: NHS, 2015. Available: www.england.nhs.uk/nhs-standard-contract/15-16/
9 Honeyford K, Cooke GS, Kinderlerer A, *et al*. Evaluating a digital sepsis alert in a London multisite Hospital network: a natural experiment using electronic health record data. *J Am Med Inform Assoc* 2020;27:274–83.
10 RLBUHT. *Our e-sepsis tool - an amazing success*. NHS Liverpool University Hospitals NHS Foundation Trust, 2019. Available: https://www.rlbuht.nhs.uk/news/our-e-sepsis-tool-an-amazing-success/
11 NHS England. Hundreds of lives saved through new tech to spot sepsis. 2019. Available: https://www.england.nhs.uk/2019/08/hundreds-of-lives-saved-through-new-tech-to-spot-sepsis/
12 NICE Short Clinical Guidelines Technical Team. *Acutely ill patients in hospital*. NIHR, 2006: 107.
13 Islam MdM, Nasrin T, Walther BA, *et al*. Prediction of sepsis patients using machine learning approach: a meta-analysis. *Computer Methods and Programs in Biomedicine* 2019;170:1–9.
14 Wong A, Otles E, Donnelly JP, *et al*. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181:1065–70.
15 Blackburn JR, Daniels R. Identification and management of sepsis. *BMJ* 2015;350:h1367.
16 NHS Standard Contact Team. *NHS standard contract 2019/20 technical guidance*. NHS England, 2019. Available: www.england.nhs.uk/wp-content/uploads/2019/03/8-NHS-Standard-Contract-Technical-Guidance-1920-v1.pdf
17 Jarvis S, Kovacs C, Briggs J, *et al*. Aggregate national early warning score (news) values are more important than high scores for a single vital signs parameter for discriminating the risk of adverse outcomes. *Resuscitation* 2015;87:75–80.
18 Warren LR, Harrison M, Arora S, *et al*. Working with patients and the public to design an electronic health record interface: a qualitative mixed-methods study. *BMC Med Inform Decis Mak* 2019;19:250.
19 Hoeksma J. Digitalhealth. 2021. Available: https://www.digitalhealth.net/2021/04/dedalus-says-ics-future-requires-open-health-eco-system/
20 Nice. *Acutely ill patients in hospital. Clinical guideline 50*. 2007: 1–107.
21 Wattanasit P, Khwannimit B. Comparison the accuracy of early warning scores with qsofa and SIRS for predicting sepsis in the emergency department. *Am J Emerg Med* 2021;46:284–8.
22 Tai HC-H, Yeh C-C, Chen Y-A, *et al*. Utilization of systemic inflammatory response syndrome criteria in predicting mortality among geriatric patients with influenza in the emergency department. *BMC Infect Dis* 2019;19:639.
23 Finkelsztein EJ, Jones DS, Ma KC, *et al*. Comparison of qsofa and SIRS for predicting adverse outcomes of patients with suspicion of sepsis outside the intensive care unit. *Crit Care* 2017;21:73.
24 McLymont N, Glover GW. Scoring systems for the characterization of sepsis and associated outcomes. *Ann Transl Med* 2016;4:527.
25 Hwang MI, Bond WF, Powell ES. Sepsis alerts in emergency departments: a systematic review of accuracy and quality measure impact. *West J Emerg Med* 2020;21:1201–10.
26 Wilson FP, Martin M, Yamamoto Y, *et al*. Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. *BMJ* 2021;372:m4786.

BMJ Health & Care Informatics

A natural language processing approach to categorise contributing factors from patient safety event reports

Azade Tabaie [1], Srijan Sengupta,[2] Zoe M Pruitt [3], Allan Fong [1]

To cite: Tabaie A, Sengupta S, Pruitt ZM, *et al*. A natural language processing approach to categorise contributing factors from patient safety event reports. *BMJ Health Care Inform* 2023;**30**:e100731. doi:10.1136/bmjhci-2022-100731

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100731).

Received 29 December 2022
Accepted 12 May 2023

[1]Center for Biostatistics, Informatics, and Data Science, MedStar Health Research Institute, Washington, District of Columbia, USA
[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA
[3]National Center for Human Factors in Healthcare, MedStar Health Research Institute, Washington, District of Columbia, USA

Correspondence to
Dr Azade Tabaie;
azade.tabaie@medstar.net

ABSTRACT

Objectives The objective of this study was to explore the use of natural language processing (NLP) algorithm to categorise contributing factors from patient safety event (PSE). Contributing factors are elements in the healthcare process (eg, communication failures) that instigate an event or allow an event to occur. Contributing factors can be used to further investigate why safety events occurred.

Methods We used 10 years of self-reported PSE reports from a multihospital healthcare system in the USA. Reports were first selected by event date. We calculated χ^2 values for each ngram in the bag-of-words then selected N ngrams with the highest χ^2 values. Then, PSE reports were filtered to only include the sentences containing the selected ngrams. Such sentences were called information-rich sentences. We compared two feature extraction techniques from free-text data: (1) baseline bag-of-words features and (2) features from information-rich sentences. Three machine learning algorithms were used to categorise five contributing factors representing sociotechnical errors: communication/hand-off failure, technology issue, policy/procedure issue, distractions/interruptions and lapse/slip. We trained 15 binary classifiers (five contributing factors * three machine learning models). The models' performances were evaluated according to the area under the precision-recall curve (AUPRC), precision, recall, and F1-score.

Results Applying the information-rich sentence selection algorithm boosted the contributing factor categorisation performance. Comparing the AUPRCs, the proposed NLP approach improved the categorisation performance of two and achieved comparable results with baseline in categorising three contributing factors.

Conclusions Information-rich sentence selection can be incorporated to extract the sentences in free-text event narratives in which the contributing factor information is embedded.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Contributing factors are important in patient safety event (PSE) report analysis, but the language associated with contributing factors could be subtle and might be embedded in other statements. This makes extracting contributing factors challenging through either manual analysis or machine learning approaches.

WHAT THIS STUDY ADDS

We explored the use of a natural language processing algorithm leveraging the unstructured PSE reports to identify information-rich sentences and demonstrated how this method improved classification performance of contributing factors in PSE reports.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

This approach can be used in near real-time to reduce the burden of manually extracting the factors influencing a patient's safety incident.

INTRODUCTION

Patient safety event (PSE) reporting systems aim to identify safety hazards by encouraging hospital staff to report on errors and potential errors in the hospital system.[1 2] Although PSE reports are limited in that they are often voluntary and only captures a small percentage of the actual prevalence of hazards, these reports have been demonstrated to still be a valuable lens to understand and improve patient safety.[3–6]

PSE reporting systems collect structured and unstructured data. The unstructured data include information about events, such as the contributing factors (CFs) relating to events and patient condition.[7] CFs are important as they represent the factors influencing patient safety incidents (eg, sociotechnical issues, communication failures, technology issues).[8–10] Although identifying and mitigating CFs could improve patient safety, the language associated with CFs could be subtle and difficult to extract as CFs might not always be explicitly described as CFs. It could be interjected between other statements, which makes extracting CFs and using current document-level natural language processing (NLP) and machine learning approaches challenging and often relies on time-intensive manual review. In the example below, while the CF distraction, there is only one sentence about a distraction:

'Registered nurse (RN) was preparing patient for left eye surgery, verified site and

procedure, consent for left eye surgery signed by the patient at bedside. RN was interrupted and inadvertently placed the eye drop for the procedure into the right eye. Patient was notified. Doctor was at bedside'.

The contributions of this work are twofold. First, we explored the use of NLP techniques to categorise CFs from PSE reports. Specifically, we investigated the utility of identifying information-rich ngrams and sentences in categorising CFs. Second, we employed three machine learning algorithms to categorise CFs: logistic regression with elastic net regularisation (elastic net), XGBoost and feed-forward neural network (FFNN).

BACKGROUND

PSE reports

PSE reports contain information regarding adverse events and errors in healthcare.[11] PSE reports contain both structured and unstructured data. For example, the relevant department and level of patient harm are reported as structured data. The event narrative is reported as unstructured, free-text data. While reporting systems encourage reporters to annotate reports with structured, easily searchable data, there are known limitations to reporting systems. They often rely on self-reporting, only captures a small per cent of hazards, can sometimes be bias based on who or what departments are reporting.[12 13] Also, the definition of taxonomies can be confusing.[14]

An example of this is the annotation or coding of CFs. Although reporting systems can give checkbox options to reporters to select associated CFs, they are used infrequently. As a result, relevant information about CFs would only occur in the free-text event narratives.

CFs in PSE

CFs are elements in the healthcare process (eg, sociotechnical issues, communication failures) that instigate an event or allow an event to occur. Human factor models, such as Systems Engineering Initiative for Patient Safety and Human Factors Analysis and Classification System were developed to categorise CFs.[15 16] These CFs can be used to understand changes that need to be made to the system or further investigate why events occurred (eg, interviews, observations).[7]

Challenges with identifying CFs

Although CFs can be selected from a predefined list by reporters, PSE reports are often recorded without a CF indicated in the structured field.[17] Instead, these factors are often described in free-text narratives. Extracting CFs from free text can be challenging because CFs are often interspersed with other text, requiring a time-consuming manual review to extract the information.

Natural language processing

NLP is an algorithmic method for extracting relevant information from free text. In this study, we hypothesise that an NLP approach that uses a sentence selection strategy will successfully identify CFs.

In generating features from free text, the standard options are to either use the term frequency-inverse document frequency (TF-IDF) matrix or word embedding techniques.[18–21] Using bag-of-words and its associated TF-IDF matrix for a categorisation task leads to a high-dimensional feature space that requires a strong computational capacity to train a model. Moreover, the less informative features can add noise to the free-text data and lead to less accurate model performance. On the other hand, while popularly used word-embedding models such as word2vec can help reduce the dimensionality of the problem, utilising word embeddings comes with a loss of interpretability since the original terms are replaced by numeric vectors.[15] Our proposed approach was inspired by previous work in identifying important sentence in free-text categorisation.[22 23] In this study, we hypothesised that an NLP approach, such as a sentence selection algorithm, could be used as a remedy by enabling noise reduction by filtering out the less informative parts of a free-text while preserving interpretability.

METHODS

We explored using an NLP approach to select information-rich sentences relating to CFs information. Then, we used three machine learning algorithms to categorise five sociotechnical CFs. Finally, the effect of the proposed methods on categorisation performance was assessed. Figure 1 demonstrates a summary of the methods used in this study. The Institutional Review Board approved this study.

Data and CF description

The self-reported PSE reports from November 2011 to October 2021 from a multihospital healthcare system in the mid-Atlantic region of the USA were included in this study. In the reporting system, reporters can select over 20 CF options from a list. The CFs are presented to the reporter as a checkbox. Reporters can select none, one or multiple CFs. For this study, we used reports with at least one CF selected by the user to have ground truth for all the included PSE reports. The list of reported CFs and a free-text brief factual description of the event were extracted for each PSE report.

Contributing factors

This study focused on five labels of reported PSE CFs associated with sociotechnical errors: communication/hand-off failure, technology issue, policy/procedure issue, distractions/interruptions and lapse/slip. These five CFs were among our data set's 10 most frequent CFs. Communication/hand-off failure refers to the problems with shift change, patient transfers and information exchange between providers. Technology issues refer to problems with health information technology and medical devices. Policy/procedure issues refer to

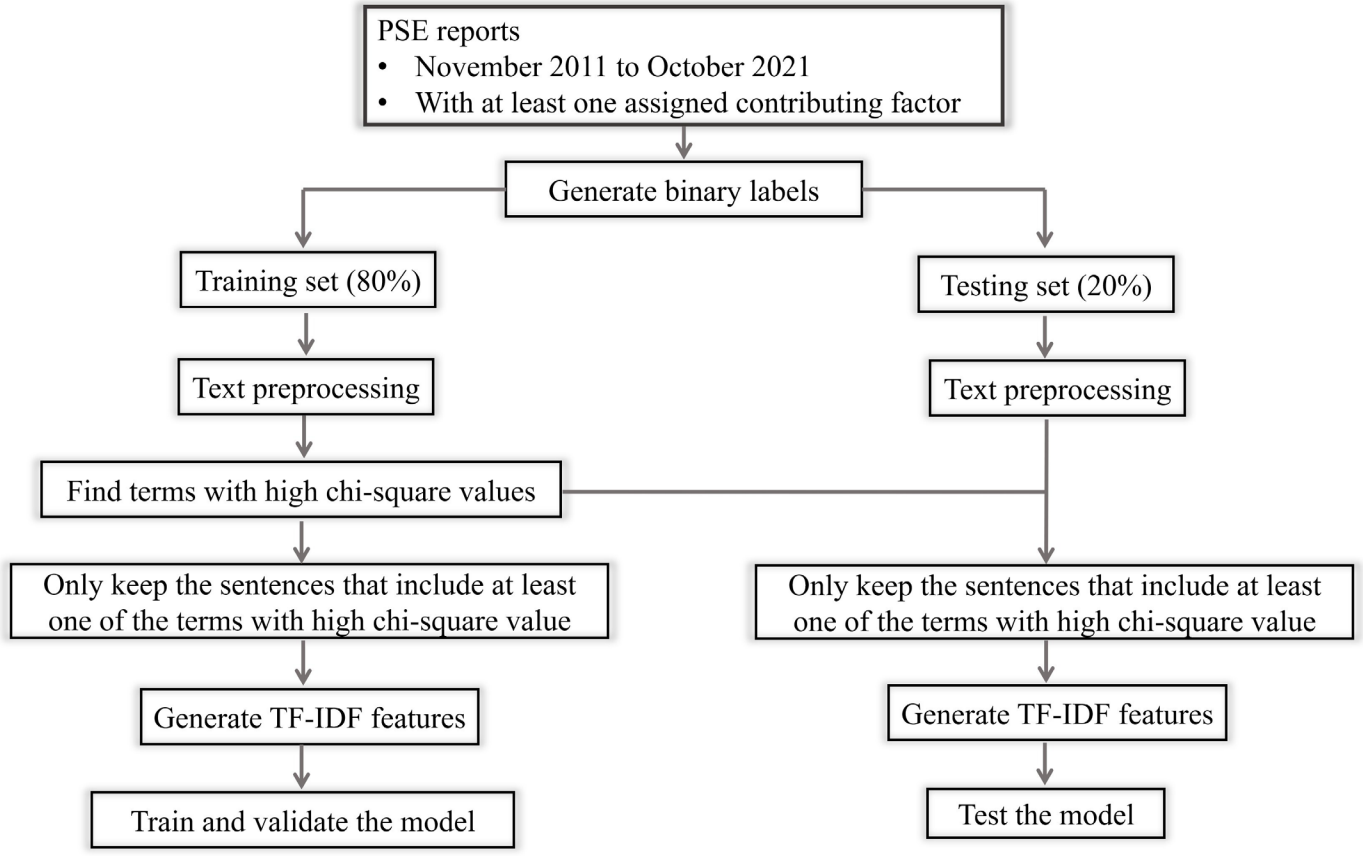

Figure 1 The summary of the methods. PSE, patient safety event; TF-IDF, term frequency-inverse document frequency.

confusing, absent or inappropriate guidelines. Distraction/interruption refers to issues when providers are diverted to a second task before completing the initial task. Lapse/slip refers to issues in human performance, such as accidentally pushing the wrong button.[24]

Text preprocessing

A PSE report can contain multiple CFs. A binary label (ie, one or zero) was assigned to each PSE report before categorising each CF. Text preprocessing is explained in online supplemental appendix A.

Selecting Information-rich terms and sentences

Our proposed NLP approach starts with identifying the information-rich ngrams in each categorisation task. We calculated χ^2 value (for every ngram that was identified in the bag-of-words free-text preprocessing step). χ^2 was calculated using the two-way contingency table of a ngram (t) and a CF (c). In Equation 1, A is the number of times t and c co-occur, B is the number of times t occurs without c, C is the number of times c occurs without t, D is the number of times neither c nor t occurs and N is the total number of PSE reports in the cohort.

$$X^2 = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

Equation 1. χ^2 calculation to identify information-rich ngrams.

χ^2 measures the degree of association between a specific ngram and the outcome label. This association can indicate a positive or negative relationship between an ngram and a CF; therefore, information-rich ngrams were the ones that achieved a higher χ^2.[25] This approach is motivated by previous work utilising χ^2.[26–28] χ^2 is computationally fast, and the results are easily interpretable.

Finally, selected sentences have at least one of the information-rich ngrams. These sentences will be referred to as 'information-rich sentences'. Concatenating all the information-rich sentences for a PSE report, we produced the free-text for generating the feature matrix.

Machine learning models

Using the bag-of-words from the preprocessed reports, we calculated the TF-IDF matrix associated with the PSE reports. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents and it is calculated by multiplying two metrics: the number of times a word appeared in a document and the inverse document frequency of the word across a set of documents. TF-IDF is a popular method to translate free-text to numerical features in training machine learning models. The data were split into a training set (80%) and a testing set (20%) using stratified sampling. All the preprocessing steps were then applied to the training set, and the same bag-of-words

was incorporated to calculate the TF-IDF matrix of the PSE reports in the testing data.

To assess the effect of the sentence selection method, we used three machine learning strategies (elastic net, XGBoost and FFNN) to categorise the PSE reports and trained separate binary categorisation models for each of the five CFs. Multiple sociotechnical CFs could be assigned to a PSE report; therefore, training a multi-label classification is possible. However, the information-rich sentence selection approach selects different sets of information-rich ngrams for each CF leading to different feature matrices for each classifier; therefore, we trained separate binary categorisation models for each CF. The top N information-rich ngrams were identified through χ^2 calculation for each categorisation task. We set N values as 2, 5, 10, 40, 60 and 100. We compared the performance of these models with their associated bag-of-words, baseline models in which no sentence selection algorithm was applied.

Elastic net

We employed a logistic regression model with elastic net regularisation, which is a weighted combination of least absolute shrinkage and selection operator (LASSO or L1) and ridge (L2) regularisations.[29] Elastic net can remove the effect of the insignificant features by setting their estimated coefficient to zero and lower the effect of the less significant features by pushing their estimated coefficient towards zero while adding more weights to the more important features. Elastic net model is easy to implement and does not require high computation power. Such characteristics make this model an accepted baseline in machine learning-based studies.[30 31] We used elastic net as the benchmark model and compared its results with more complex categorisation methods.

XGBoost

This model is a decision tree-based boosting ensemble machine learning algorithm.[32] In a boosting algorithm, many weak learners are trained to correctly categorise the observations incorrectly classified in the previous training rounds. XGBoost uses a shallow tree as a weak learner and proved to have a decent performance in the case of class-imbalanced data classification.[31 33]

Feed-forward neural network

The feed-forward model is a simple form of a neural network as information is only processed in one direction, and the connection between nodes does not form a cycle.[34] The main benefit of this model is that FFNN accounts for higher order interactions among the input features.[31] We used one hidden layer, a binary cross-entropy loss function and Adam optimiser to train this model.[35]

Instead of data-level solutions (Synthetic Minority Oversampling TEchnique (SMOTE),[36] under-sampling, oversampling), we incorporated algorithm-level solutions (eg, boosting methods, neural networks) as remedy to the class-imabalanced problem in this categorisation problem. To evaluate the performance of the trained models, we calculated area under the operative characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV) or precision, negative categorisation value (NPV), accuracy, F-1 score and area under the precision-recall curve (AUPRC). Since the models were trained on class-imbalanced data, we focused on the AUPRC values to identify the best-performing models.

RESULTS

Descriptive summary

Of 70 680 self-reported PSE reports from November 2011 to October 2021 were extracted from PSRS. The PSE reports with unknown CFs were excluded from the study. In total, 53 899 PSE reports met the inclusion criterion. Online supplemental appendix B presents the frequency of the five CFs in our cohort. Not all CFs were included in this analysis; therefore, the percentages in online supplemental appendix B do not add up to one.

The PSE reports data were deidentified in terms of patient's name, date of birth, etc. However, the event narratives were not deidentified. The data are stored behind our Healthcare System's firewall, and it is not accessible to unauthorised users.

Information-rich terms and sentences

Table 1 presents the ngrams that were most influential when categorising the specific CFs. These ngrams were associated with high χ^2 and represented information-rich ngrams in each categorisation task.

The 25th percentile, median and 75th percentile of the number of sentences per PSE report were 4, 6, and 9, respectively. Communication/hand-off failure was the most sensitive and lapse/slip was the least sensitive to the changes in the number of selected information-rich ngrams.

Filtering out less relevant text changes the number of features to incorporate in a categorisation model. Figure 2 presents how the machine learning models' input dimension changed as we increased the number of selected information-rich ngrams. The input dimension would be almost 6700 if no sentence selection was applied. Unsurprisingly, the input dimension increased with the number of selected information-rich ngrams. Lapse/slip had the largest difference, while distractions/interruptions had the smallest difference.

Model comparison

The average number of sentences associated with each CF before and after incorporating information-rich sentence selection algorithm is presented in online supplemental appendix C. Moreover, the number of PSE reports grouped by sociotechnical CFs in training and testing data sets are included in online supplemental appendix D.

The XGBoost model performed best in categorising technology issues with AUPRC of 0.56 (figure 3). The

Table 1 Top five information-rich ngrams identified through χ^2 feature selection in each categorisation task

	Ngrams	χ^2 value
Communication/handoff failure	fall nurs order emerg depart depart	3786.7 2793.9 2462.0 1902.3 1853.6
Policy/procedure issue	min later fall polici glucomet glucose glucos run	1714.5 1474.9 1429.4 1408.1 1373.5
Technology issue	cgi field dose system medconnect	4917.1 1664.5 1646.7 1488.2 1425.2
Lapse/slip	dose pharmacy order pharmacist med	2158.7 1520.0 929.4 901.2 664.2
Distractions/interruptions	data entri vaccin error tablet medic	1686.5 1467.1 1452.1 1437.0 895.2

The words are stemmed.

FFNN model outperformed the other two models in categorising policy/procedure issues with AUPRC of 0.66. The XGBoost, elastic net and FFNN models achieved comparable performance in categorising communication/hand-off failure (AUPRCs=0.6, 0.58, 0.6), lapse/slip (AUPRCs=0.17, 0.15, 0.18) and distractions/interruptions (AUPRCs=0.24, 0.24, 0.22).

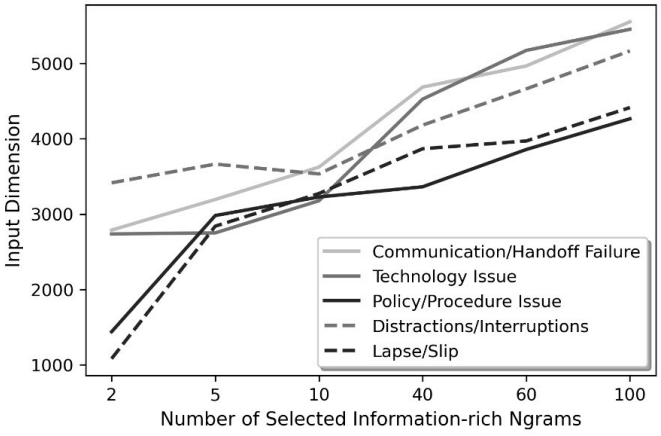

Figure 2 Input dimension vs the number of selected information-rich ngrams. Selecting the information-rich sentences, which include at least one of the selected information-rich ngrams, led to a new input feature dimension for contributing factor categorisation.

Besides AUPRC values, AUROC, sensitivity, specificity, PPV, NPV, F-1 score and accuracy were calculated for each model, and the results are included in the online supplemental appendix D. The elastic net model achieved highest AUROC (0.948) with baseline model in categorising distractions/interruption, highest sensitivity (0.879) with 60 information-rich ngrams in categorising distractions/interruptions, highest specificity (0.984) with two information-rich ngrams in categorising policy/procedure issue, highest PPV (0.692) with two information-rich ngrams in categorising technology issue, highest NPV (0.996) with 60 information-rich ngrams in categorising distractions/interruptions and highest accuracy (0.692) in categorising technology issue. The FFNN model obtained the highest F-1 score (0.962) with five information-rich ngrams in categorising policy/procedure issue.

DISCUSSION

This study used an NLP approach to identify information-rich sentences to categorise five CFs associated with sociotechnical errors. Automating the identification of CFs may help safety officers identify the CFs leading to safety issues in their organisations.

Utility of Information-rich sentences

Working with noise in the data is one of the challenges when dealing with free-text formatted input data. Filtering the input sentences and selecting more informative ones work as a solution to reduce the noise in the data when dealing with free-text categorisation of CFs. Finding a balance between removing the noise and keeping a sufficient number of features to have a well-trained model is a critical task.

The information-rich ngrams were selected according to their χ^2 values, indicating the association between ngrams and a CF. The most relevant term for communication/hand-off failure in our cohort was *fall* followed by *nurse* and *order*. *Glucose* readings were repeatedly identified as an important contributor to policy/procedure issues. Our data's selected information-rich ngrams for technology issues present the same trend through identifying *medconnect* and *system* as the most relevant ngrams for this CF. Besides, CGI, the stemmed form of Centigray (CGY), was the top information-rich ngram for technology issues. CGY is a measurement of absorbed radiation. This result suggests for technology-related CFs associated with radiation treatments such as with cyberknife radiation treatments. The ngrams with high χ^2 values associated with lapse/slip were *dose*, *pharmacy* and *order*. The pattern among these ngrams indicates that the medication-related tasks were more prone to be affected by this CF. Our analysis indicated that the data entry process was affected by distractions/interruptions as some of the information-rich ngrams for this CF were *data entry* and *error*. Box 1 presents examples of information-rich ngrams, which identified information-rich sentences in PSE reports.

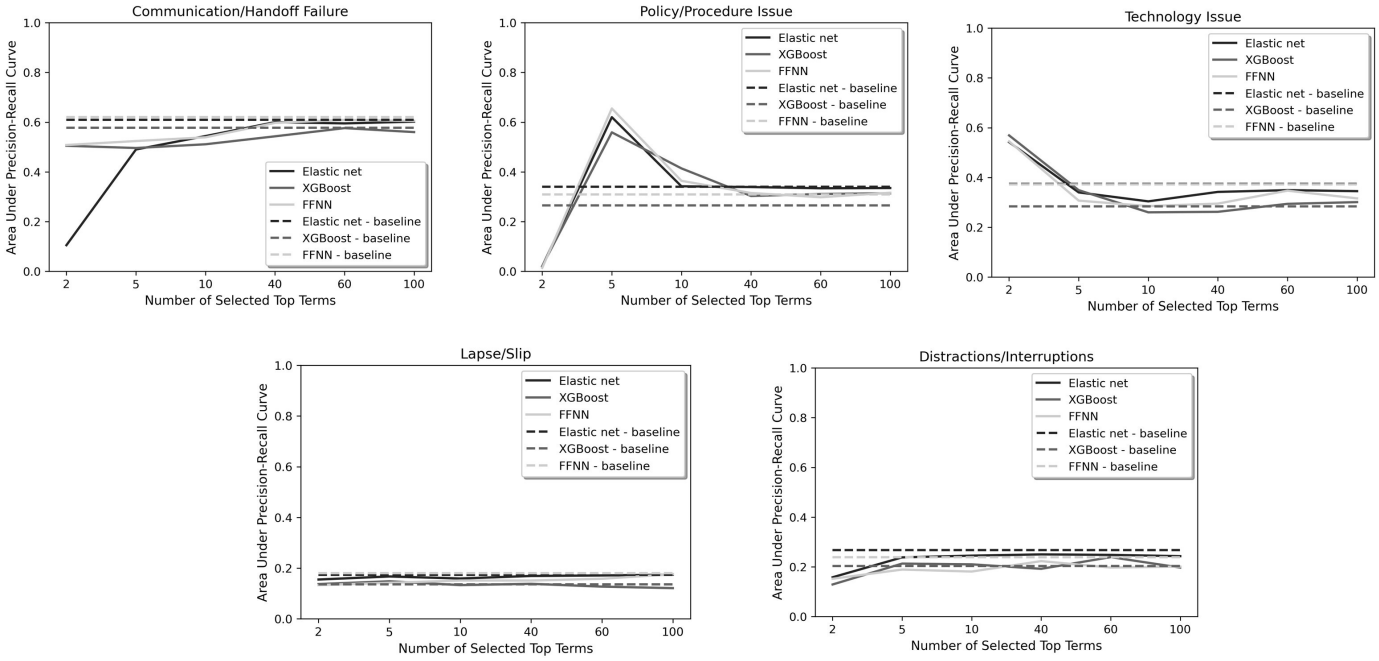

Figure 3 Area under precision-recall trends. For each of the five contributing factors, this figure presents how the value of AUPRC score changes as we only include the information-rich sentences in PSE reports containing the information-rich ngrams. AUPRC, area under the precision-recall curve; FFNN, feed-forward neural network; PSE, patient safety event.

We included unigrams and bigrams in the bag-of-words to calculate their χ^2 value. Information-rich unigrams were more common than bigrams, perhaps because unigrams contain more generalisable information. However, bigrams can convey more specific information, but they are susceptible to noise. Further investigation is needed to assess the effect of bigrams in the information-rich sentence selection algorithm.

Performance boosting

The proposed method proved its benefit by improving the results of categorising two CFs and achieving comparable results with the baseline models for three CFs in this analysis. Figure 2 shows a near-linear relationship between the number of input dimensions and the selected information-rich ngrams with higher χ^2 values. However, the AUPRC plots in figure 3 indicate that incorporating 2, 5 or 10 information-rich ngrams in selecting the information-rich sentences for the model training process may lead to better performance metrics. Thus, it can be inferred that the balance between removing noise and preserving features for an accurate model can be obtained by selecting the sentences containing the top 5 or 10 information-rich ngrams with the highest χ^2 values. This balance depends on the information embedded in the unstructured data.

Content depending

The elastic net model did not perform remarkably better than the neural network or ensemble model, implying that including the higher order interaction between the terms improved the categorisation performance. Identifying bigrams as information-rich terms also indicates the importance of the interactions among the features. FFNN and XGBoost models achieved comparable results. The difference between the three models' performance was negligible when categorising the CFs with lower prevalence, such as lapse/slip and distraction/interruption.

The information-rich sentence selection algorithm improved the performance of categorising two CFs, policy/procedure and technology issues. The performance improvement may also indicate that healthcare providers tend to be more consistent in their language to record safety events related to policy/procedure and technology issues. Distraction/interruption and lapse/slip had relatively smaller sample size compared with the other three CFs. However, the effect of information-rich sentence selection boosted the performance of distraction/interruption better. This is also an indication of having consistent language in recording safety incidents related to distraction/interruption.

Information-rich sentence selection is a data-driven approach; therefore, depending on the context of the unstructured input, the outcomes of using this approach could be different. For instance, the information-rich sentence selection approach improved the AUPRC values of all three machine learning categorisation models compared with baseline models in categorising policy/procedure and technology issue incorporating 5 and 2 information-rich ngrams, respectively. While the performance was not boosted for communication/hand-off failure, applying information-rich sentence selection approach using 100 information-rich terms led to similar results compared with using the entire PSE report for the categorisation task. Equivalently, using the approach

Box 1 Instances of information-rich sentence selection algorithm applied on patient safety event (PSE) reports from different contributing factor categories.

Contributing factor
Communication/hand-off failure.
PSE report (brief factual description)
Patient was taken to nuclear medicine via transport for a scheduled stress test. Once he got to NM, the test was cancelled. Patient had drunk coffee with his breakfast because there was no NPO order in place for the test.
Information-rich Ngram
Selected information-rich sentence
Patient had drunk coffee with his breakfast because there was no NPO order in place for the test.
Contributing factor
Policy/procedure issue
PSE report (brief factual description)
A glucose test was performed at (time stamp 1) on patient by (nures 1) with a result of 36 mg/dL. The test was performed at (time stamp 2) by (nurse 2) with a result of 139 mg/dL, which was 1 hour and 4 min later. The Hypoglycaemia Policy states that a patient with a glucose less than 40 mg/dL should be treated and a glucose run every 15 min until the glucose returns to 90 mg/dL.
Information-rich Ngram
Selected information-rich sentence
A glucose test was performed at (time stamp 1) on patient by (nures 1) with a result of 36 mg/dL. The Hypoglycaemia Policy states that a patient with a glucose less than 40 mg/dL should be treated and a glucose run every 15 min until the glucose returns to 90 mg/dL.
Contributing factor
Technology issue.
PSE report (brief factual description)
I was unable to gain access to pyxis. Rebooted system and tried several interventions but unsuccessful. ICU and ED called to report inability to gain access to pyxis. Carefusion called and stated that the database was disconnected from the system and unable to diagnose problem at this time. Instructed to call help line and high priority ticket initiated. Patient began seizing. Medication system down and unable to obtain ativan in the ED. Nurse had to physically go to the pharmacy to obtain medicine.
Information-rich Ngram
Selected information-rich sentence
Rebooted system and tried several interventions but unsuccessful. Carefusion called and stated that the database was disconnected from the system and unable to diagnose problem at this time. Medication system down and unable to obtain ativan in the ED.
Contributing factor
Lapse/slip.
PSE report (brief factual description)
Order for an HIV med entered on the wrong patient. The pharmacists did not quesiton why the patient was ordered for only one HIV medication. The doctor called one afternoon asking how did this mistake happen and not be caught. At that time, that is when the pharmacists was made aware of the mistake.
Information-rich Ngram
Pharmacist.
Selected information-rich sentence
The pharmacists did not quesiton why the patient was ordered for only one HIV medication. At that time, that is when the pharmacists was made aware of the mistake.

Continued

Box 1 Continued

Contributing factor
Distractions/interruptions.
PSE report (brief factual description)
Three prescriprions were e-scribed for one of our long-term patients here at store #N. All prescriptions were prepared and dispensed expediciously since our client was in a hurry to make his ride. All the medications were controlled except for one medication. The next day, we received a call from the doctor's office, which happens to be a first time doctor for this client, stating that one of the medications were to be dispensed at a later date on ((date)). Unfortunately, missed that date at data entry as I performed the data entry of the prescriptions. The team did contact the patient and informed him of the prescriber's specific directions in regards to that one prescription.
Information-rich Ngram
Selected information-rich sentence
Unfortunately, missed that date at data entry as I performed the data entry of the prescriptions.

with 100 information-rich ngrams for lapse/slip, and 60 information-rich ngrams for distraction/interruption did not improve the baseline AUPRC but achieved similar results with a filtered data and lower dimensional input for the machine learning models.

Limitations

Our study has limitations. First, our data came from a single health system and may reflect the specific language to the system. While PSE reports are recorded across all healthcare systems, the external application of our methods on data from other facilities may be biased. Second, our models were developed and evaluated based on a retrospective cohort; therefore, the performance may deteriorate when the method is applied to real-time data. Third, the five CFs included in this study are not the only CFs representing sociotechnical error. We explored the use of the NLP approach on only five sociotechnical CFs. This approach can be explored to categorising the rest of the sociotechnical CFs. Fourth, although we investigated the results and provided insights into the models' decision-making process, our study did not benefit from human factor expert input and critical analysis. Fifth, we selected TF-IDF, a widely known text feature extraction technique, and did not examine all text feature extraction methods (eg, YAKE!, rake, etc). Sixth, the CFs used in this study were assigned to PSE reports by reporters of safety incidents. The human-selected CFs could introduce some level of uncertainty to the labels. Seventh, we tested six values for the number of information-rich ngrams (ie, 2, 5, 10, 40, 60 and 100). Other values could be incorporated to measure the advantage of employing information-rich sentence selection algorithm. Finally, we excluded PSE reports, which did not have assigned CF, that can affect the performance of the models in near-real time applications.

Identifying and addressing CFs is critical for improving patient safety as these, often latent, themes are prevalent

across departments, event types and service lines. Being able to more readily identify CFs across departments, event types and service lines can provide patient safety leaders and healthcare systems awareness and insights to address safety events and hazards more at a system level.[37 38] This analysis is limited to what gets reported. While this is a useful start and one lens to understand CFs, a broader multiperspective approach is needed to understand all dimensions of CFs, including from the patient's perspective.[12 13] When analysing a large body of PSE reports and the associated CFs, it is essential to consider potential language bias and department bias (ie, using reports to 'blame and shame' other departments) in the recorded data.

CONCLUSION

We explored an NLP approach to categorise five socio-technical CFs in PSE reports. We demonstrated the utility of information-rich sentence selection in free-text categorisation. This approach can be used in near real-time to reduce the burden of manually extracting the factors influencing a patient's safety incident. Information such as patient feedback and complaints can be paired with the findings of this study to inform strategies around patient safety efforts and help teams make decisions.

Contributors Guarantor: AT, AF. Conceived study design: AT, AF. Contributed to data analysis: AT, AF. Contribute to visualisation: AT. Wrote the manuscript: AT, SS, ZMP, AF. Reviewed and edited manuscript: AT, SS, ZMP, AF.

Funding This work was supported by the National Institutes of Health grant number 1R01LM013309-01.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iDs

Azade Tabaie http://orcid.org/0000-0003-1869-5923
Zoe M Pruitt http://orcid.org/0000-0002-2749-9652
Allan Fong http://orcid.org/0000-0002-7550-1569

REFERENCES

1 Fong A. Realizing the power of text mining and natural language processing for analyzing patient safety event narratives: the challenges and path forward. *J Patient Saf* 2021;17:e834–6.
2 Archer S, Hull L, Soukup T, *et al*. Development of a theoretical framework of factors affecting patient safety incident reporting: a theoretical review of the literature. *BMJ Open* 2017;7:e017155.
3 Mitchell I, Schuster A, Smith K, *et al*. Patient safety incident reporting: a qualitative study of thoughts and perceptions of experts 15 years after 'to err is human. *BMJ Qual Saf* 2016;25:92–9.
4 Macrae Carl. The problem with incident reporting. *BMJ Qual Saf* 2016;25:71–5.
5 Roehr B. US hospital incident reporting systems do not capture most adverse events. *BMJ* 2012;344(jan13 2):bmj.e386.
6 Flynn EA, Barker KN, Pepper GA, *et al*. Comparison of methods for detecting medication errors in 36 hospitals and skilled-nursing facilities. *Am J Health Syst Pharm* 2002;59:436–46.
7 Pronovost PJ, Thompson DA, Holzmueller CG, *et al*. Toward learning from patient safety reporting systems. *Journal of Critical Care* 2006;21:305–15. 10.1016/j.jcrc.2006.07.001 Available: https://doi.org/10.1016/j.jcrc.2006.07.001
8 Holmström AR, Järvinen R, Laaksonen R, *et al*. Inter-Rater reliability of medication error classification in a voluntary patient safety incident reporting system Haipro in Finland. *Research in Social and Administrative Pharmacy* 2019;15:864–72.
9 Amaniyan S, Faldaas BO, Logan PA, *et al*. Learning from patient safety incidents in the emergency Department: a systematic review. *J Emerg Med* 2020;58:234–44.
10 Lacson R, Cochon L, Ip I, *et al*. Classifying safety events related to diagnostic imaging from a safety reporting system using a human factors framework. *Journal of the American College of Radiology* 2019;16:282–8.
11 Lawton R, McEachan RRC, Giles SJ, *et al*. Development of an evidence-based framework of factors contributing to patient safety incidents in hospital settings: a systematic review. *BMJ Qual Saf* 2012;21:369–80.
12 Puthumana JS, Fong A, Blumenthal J, *et al*. Making patient safety event data actionable: understanding patient safety analyst needs. *J Patient Saf* 2021;17:e509–14.
13 Pronovost P, Morlock LL, Sexton B. Improving the value of patient safety reporting systems. In: *Advances in patient safety: New directions and alternative approaches. Vol 1. Assessment*. Rockville, MD: Agency for Healthcare Research and Quality, 2008.
14 Wood KE, Nash DB. Mandatory state-based error-reporting systems: current and future prospects. *Am J Med Qual* 2005;20:297–303.
15 Holden RJ, Carayon P, Gurses AP, *et al*. SEIPS 2.0: a human factors framework for studying and improving the work of Healthcare professionals and patients. *Ergonomics* 2013;56:1669–86.
16 Diller T, Helmrich G, Dunning S, *et al*. The human factors analysis classification system (HFACS) applied to health care. *Am J Med Qual* 2014;29:181–90.
17 Magrabi F, Ong MS, Runciman W, *et al*. An analysis of computer-related patient safety incidents to inform the development of a classification. *J Am Med Inform Assoc* 2010;17:663–70.
18 Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*. Stroudsburg, PA, USA, n.d.: 1532–43.
19 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their Compositionality. *Adv Neural Inf Process Syst* 2013;26.
20 Peters M, Neumann M, Zettlemoyer L, *et al*. Dissecting contextual word embeddings: architecture and representation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Stroudsburg, PA, USA, Brussels, Belgium.
21 Devlin J, Chang MW, Lee K, *et al*. Bert: pre-training of deep Bidirectional transformers for language understanding. 2018.
22 Ko Y, Park J, Seo J. Improving text Categorization using the importance of sentences. *Information Processing & Management* 2004;40:65–79.
23 Ogura Y, Kobayashi I. Text classification based on the latent topics of important sentences extracted by the Pagerank algorithm. In: *Proceedings of the conference. Association for Computational Linguistics. Meeting*. n.d.: 46–51.
24 Wright J, Lawton R, O'Hara J, *et al*. Assessing risk: a systematic review of factors contributing to patient safety incidents in hospital settings. improving patient safety through the involvement of patients: development and evaluation of novel interventions to engage patients in preventing patient safety incidents and protecting them against unintended harm. October 2016.

25 Chen YT, Chen MC. Using Chi-square Statistics to measure similarities for text Categorization. *Expert Systems with Applications* 2011;38:3085–90. 10.1016/j.eswa.2010.08.100 Available: https://doi.org/10.1016/j.eswa.2010.08.100
26 Fothergill R, Cook P, Baldwin T. Evaluating a topic Modelling approach to measuring corpus similarity. In: *Int Conf Lang Resour Eval*. 2016: 273–9.
27 Kilgarriff A. Using word frequency lists to measure corpus homogeneity and similarity between Corpora. *InFifth Workshop on Very Large Corpora* 1997.
28 Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical Taxonomy. 20, 1997.
29 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* 2005;67:301–20.
30 Lee SJ, Weinberg BD, Gore A, *et al*. A Scalable natural language processing for Inferring BT-RADS Categorization from unstructured brain magnetic resonance reports. *J Digit Imaging* 2020;33:1393–400.
31 Pfob A, Sidey-Gibbons C, Lee H-B, *et al*. Identification of breast cancer patients with pathologic complete response in the breast after Neoadjuvant systemic treatment by an intelligent vacuum-assisted biopsy. *Eur J Cancer* 2021;143:134–46.
32 Chen T, Guestrin C. Xgboost: A Scalable tree boosting system. *KDD* 2016;13:785–94.
33 Tabaie A, Orenstein EW, Nemati S, *et al*. Predicting presumed serious infection among hospitalized children on central venous lines with machine learning. *Comput Biol Med* 2021;132:104289.
34 Svozil D, Kvasnicka V, Pospichal J. Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems* 1997;39:43–62.
35 Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014.
36 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: synthetic minority over-sampling technique. *Jair* 2002;16:321–57.
37 Clarke JR. How a system for reporting medical errors can and cannot improve patient safety. *Am Surg* 2006;72:1088–91;
38 Howell AM, Burns EM, Hull L, *et al*. International recommendations for national patient safety incident reporting systems: an expert Delphi consensus-building process. *BMJ Qual Saf* 2017;26:150–63.

BMJ Health & Care Informatics

Measures of socioeconomic advantage are not independent predictors of support for healthcare AI: subgroup analysis of a national Australian survey

Emma Kellie Frost ,[1] Pauline O'Shaughnessy,[2] David Steel,[3] Annette Braunack-Mayer,[1] Yves Saint James Aquino,[1] Stacy M Carter[1]

To cite: Frost EK, O'Shaughnessy P, Steel D, *et al*. Measures of socioeconomic advantage are not independent predictors of support for healthcare AI: subgroup analysis of a national Australian survey. *BMJ Health Care Inform* 2023;**30**:e100714. doi:10.1136/bmjhci-2022-100714

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100714).

Received 04 January 2023
Accepted 11 May 2023

For numbered affiliations see end of article.

Correspondence to
Emma Kellie Frost;
emmaf@uow.edu.au

ABSTRACT

Objectives: Applications of artificial intelligence (AI) have the potential to improve aspects of healthcare. However, studies have shown that healthcare AI algorithms also have the potential to perpetuate existing inequities in healthcare, performing less effectively for marginalised populations. Studies on public attitudes towards AI outside of the healthcare field have tended to show higher levels of support for AI among socioeconomically advantaged groups that are less likely to be sufferers of algorithmic harms. We aimed to examine the sociodemographic predictors of support for scenarios related to healthcare AI.

Methods: The Australian Values and Attitudes toward AI survey was conducted in March 2020 to assess Australians' attitudes towards AI in healthcare. An innovative weighting methodology involved weighting a non-probability web-based panel against results from a shorter omnibus survey distributed to a representative sample of Australians. We used multinomial logistic regression to examine the relationship between support for AI and a suite of sociodemographic variables in various healthcare scenarios.

Results: Where support for AI was predicted by measures of socioeconomic advantage such as education, household income and Socio-Economic Indexes for Areas index, the same variables were not predictors of support for the healthcare AI scenarios presented. Variables associated with support for healthcare AI included being male, having computer science or programming experience and being aged between 18 and 34 years. Other Australian studies suggest that these groups may have a higher level of perceived familiarity with AI.

Conclusion: Our findings suggest that while support for AI in general is predicted by indicators of social advantage, these same indicators do not predict support for healthcare AI.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence (AI) has the potential to perpetuate existing biases in healthcare data sets, which may be more harmful for marginalised populations. Support for the development of AI tends to be higher among more socioeconomically privileged groups.

WHAT THIS STUDY ADDS

⇒ While general support for the development of AI was higher among socioeconomically privileged groups, support for the development of healthcare AI was not. Groups that were more likely to support healthcare AI were males, those with computer science experience and younger people.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Healthcare AI is becoming more relevant for the public as new applications are developed and implemented. Understanding how public attitudes differ among sociodemographic subgroups is important for future governance of healthcare AI.

BACKGROUND

There are currently many applications for healthcare artificial intelligence (HCAI) in various stages of development and implementation.[1] Defined as technologies that allow computer programs to perform tasks and solve problems without explicit human guidance,[2] HCAI-based systems employ algorithms to complete the tasks typically performed by health professionals. Algorithms have been trained to read ECGs,[3] detect skin cancer from smartphone images[4] and predict people's risk of disease using large-scale national data sets[5] with ostensibly comparable accuracy to current approaches.

While these technologies have the potential to improve aspects of healthcare, they also have the potential to cause harm to patients.[6] Algorithmic harms are exacerbated in already marginalised populations,[7 8] as the causes and effects of historical structural disadvantage are embedded in healthcare data sets, and training sets often exclude marginalised groups. Obermeyer *et al*[9] audited an algorithm used in the USA for determining whether patients should be referred to high-risk care, and found that patients who identified as black were less likely to be flagged by

the algorithm as needing high-risk care, despite having more comorbidities than non-black-identifying counterparts. Similarly, Seyyed-Kalantari *et al*,[8] using data from the USA, found that women, people aged under 20, those with lower socioeconomic status and black or Hispanic-identifying people were less likely to be diagnosed correctly by a chest radiograph algorithm. Factors preventing marginalised groups from accessing care in the past exist implicitly in many healthcare data sets, and algorithms trained on these data sets perpetuate these inequities.[9]

Surveys examining public attitudes towards artificial intelligence (AI) have found that certain sociodemographic characteristics are associated with higher levels of support for AI. Zhang and Dafoe[10] in a survey in the USA found that younger people, males, those with computer science experience and those with a high annual household income were more likely to be supportive of the development of AI. A survey study in the Netherlands, using a representative panel of the Dutch population, studied trust in HCAI and found that the sociodemographic characteristics associated with higher levels of trust were being male, having a higher level of education, being employed or a student and having not stayed in hospital in the past 12 months.[11] It is suggested that those who are less likely to suffer from the negative impacts of AI are more supportive of its implementation.[10–12]

We conducted a survey to examine whether Australians' attitudes towards HCAI vary with different sociodemographic characteristics.

METHOD

Our aims for this study were threefold. We aimed to (1) examine the sociodemographic variables associated with support for AI in Australia, (2) examine the sociodemographic variables associated with support for HCAI and (3) determine whether sociodemographic characteristics were associated with different preferences in AI-integrated healthcare.

This paper reports results from an analysis of the Australian Values and Attitudes toward AI (AVA-AI) survey. The survey was conducted with the Social Research Centre's *Life in Australia* (LIA) study, which regularly engages a representative panel of Australians in independent surveys.[13] A shortened version of the AVA-AI questionnaire was included in the 36th wave of the LIA study, disseminated in March 2020. The full version of the questionnaire was disseminated to a non-probabilistically sampled online panel. We used the shortened version of the questionnaire as a reference survey to produce weights for the non-probability sample that account for characteristics that influence people's propensity to participate in the online panel. A more detailed description of the data collection and weighting methodology is provided in Isbanner *et al*'s study.[14] For this analysis, we report on results from the weighted non-probability sample using data obtained from the full questionnaire.

Predictor variables

We selected predictor variables analogous to two other surveys on public attitudes towards AI: Zhang and Dafoe's study in the USA[10] and Selwyn and colleagues' study in Australia.[15] Variables used in the analysis included age group, gender, self-identification as having a chronic health condition or disability, living in a capital city, highest level of educational attainment, area of socioeconomic advantage (henceforth referred to as Socio-Economic Indexes for Areas (SEIFA)) (This study used the Australian Bureau of Statistics' SEIFA to measure the relative advantage and disadvantage of areas.[16] Participants were classified into quintiles based on the SEIFA of their area (ie, postcode) of residence, with those in quintiles 4 and 5 coded as 'least socioeconomic disadvantage', those in quintiles 2 and 3 coded as 'moderate disadvantage' and those in quintile 1 coded as 'most socioeconomic disadvantage'), household income, computer science or programming experience and speaking a language other than English at home. Additionally, we included self-reported health status as a predictor variable because evidence elsewhere indicated that health-related metrics were associated with attitudes towards HCAI.[11] A copy of the questionnaire is provided in online supplemental file 1.

We removed any responses where the participant had not responded to all predictor and outcome variables (n=17). One participant identified with a gender outside of the male/female binary. This response was removed,[17] and the limitations of this will be discussed further below. n=1983 responses were analysed.

We calculated Spearman's r coefficients to identify multicollinearity between predictor variables (table 1). Some pairs of variables were moderately correlated. Those with high self-reported health status were less likely to identify as having a disability, and those living in a capital city were more likely to live in postcodes with less socioeconomic disadvantage. We deemed these moderate correlations unlikely to have a detrimental effect on model fitting or interpretation.

Outcome variables

Eleven outcome variables were selected for the three aims of the study (table 2). Item 1 replicated a question from Zhang and Dafoe's study,[10] asking participants to indicate their level of support for the development of AI on a 5-point semantic scale from *strongly oppose* to *strongly support*. Item 2 was a question that asked participants to consider their support for HCAI in a scenario where an unexplainable algorithm was being used to analyse patient health records and suggest treatments. Item 3 asked participants to consider their support for an algorithm that diagnosed diseases more accurately than physicians but required patients to share their health record. Item 5 asked participants to consider their support for HCAI in a scenario where its development leads to physicians becoming less skilled at tasks that were replaced by

Table 1 Correlation matrix of predictor variables (Spearman's r coefficients)

Self-reported disability	0.04								
Age group	–0.16	0.23							
Education	0.26	–0.12	–0.22						
Gender	0.13	0.04	0.11	0.05					
Household income	0.07	–0.22	–0.17	0.24	0.04				
Speaks languages other than English at home	0.20	–0.13	–0.23	0.24	0.00	0.02			
Living in a capital city	0.08	–0.14	–0.19	0.18	–0.01	0.12	0.22		
SEIFA	0.05	–0.14	–0.06	0.15	–0.01	0.11	0.08	0.32	
Self-reported health status	0.08	–0.34	–0.23	0.14	–0.02	0.17	0.10	0.08	0.10
	Computer science experience	Self-reported disability	Age group	Education	Gender	Household income	Languages other than English	Living in a capital city	SEIFA index

0 indicates no correlation. Coefficients closer to 1 or –1 indicate stronger positive and negative correlations, respectively.
SEIFA, Socio-Economic Indexes for Areas.

AI. Each of these questions asked participants to indicate their level of support on a 5-point scale.

Items 5–11 were preceded by a scenario asking participants to imagine a situation where an algorithm was reading a medical test, diagnosing them with a disease and recommending treatments. Participants were asked to consider the importance of (5) explainability, (6) speed, (7) accuracy, (8) human oversight, (9) accountability, (10) cost to the healthcare system and (11) equity. Participants responded on a 5-point scale from *not at all important* to *very important*. Each outcome variable was recoded to binary categories, where the two highest categories (ie, strongly support and somewhat support, very important and extremely important) were recoded to 1 and remaining categories were coded to 0.

Statistical analysis

We generated frequency tables that incorporated the survey weights using the *questionr* package.[18] We fit separate multiple logistic regression models for each of the outcome variables using the same suite of sociodemographic variables as predictors for each. All analyses were conducted in R.[19] The *survey* package[20] was used to incorporate survey weights in the analysis and calculation of SEs. ORs are reported with accompanying p values

Table 2 Aims and outcome variables

Aim	Items used	Predictor variables
Aim 1: examine the sociodemographic variables associated with support for AI in Australia	1. Level of support for the development of AI (B01)	► Gender (A03) ► Age (A01/A02) ► Self-identifies as having a chronic health condition or disability (F18) ► Education (F07) ► Household income (F06) ► Speaks languages other than English at home (F16) ► Resides in a capital city (A04)* ► SEIFA (A04)* ► Self-reported health (F17) ► Computer science or programming experience (F05)
Aim 2: examine the sociodemographic variables associated with support for HCAI	2. Level of support for HCAI that is unexplainable (C03) 3. Level of support for HCAI that requires sharing personal data (C04) 4. Level of support for HCAI that leads to clinician deskilling (C05)	
Aim 3: determine whether sociodemographic characteristics were associated with different preferences in AI-integrated healthcare	5. Importance of explainability (C01a) 6. Importance of getting an answer quickly (C01b) 7. Importance of getting an accurate answer (C01c) 8. Importance of being able to talk to a person about one's health (C01d) 9. Importance of knowing who is responsible for one's care (C01e) 10. Importance of reducing health system costs (C01f) 11. Importance of knowing the system treats everyone fairly (C01g)	

*Residing in a capital city and SEIFA are derived from self-reported postcode.
AI, artificial intelligence; HCAI, healthcare artificial intelligence; SEIFA, Socio-Economic Indexes for Areas.

Table 3 Weighted and unweighted sample demographics

	Unweighted		Weighted	
	n	%	n	%
Computer science or programming experience				
No	1598	85.0	1603.4	85.3
Yes	281	15.0	275.6	14.7
Has chronic health condition or disability				
No	1361	72.4	1457.2	77.6
Yes	518	27.6	421.8	22.4
Age group				
18–34	572	30.4	593.2	31.6
35–54	630	33.5	640.5	34.1
55+	677	36.0	645.3	34.3
Highest level of educational attainment				
High school	603	32.1	632.3	33.7
Trade certificate/diploma	630	33.5	709.3	37.7
Bachelor's degree	452	24.1	374.7	19.9
Postgraduate degree	194	10.3	162.7	8.7
Gender				
Female	947	50.4	968.1	51.5
Male	932	49.6	910.9	48.5
Household income (per week)				
<$500	361	19.2	340.2	18.1
$500–$1999	1095	58.3	1051.5	56.0
$2000+	423	22.5	487.3	25.9
Speaks languages other than English at home				
No	1598	85.0	1473.8	78.4
Yes	281	15.0	405.2	21.6
Lives in capital city				
No	626	33.3	626.7	33.4
Yes	1253	66.7	1252.3	66.6
SEIFA				
Most disadvantage	281	15.0	294.9	15.7
Moderate	1185	63.1	1160.2	61.7
Least disadvantage	413	22.0	423.8	22.6
Self-reported health				
Excellent/very good	735	39.1	1015.1	54.0
Good/fair/poor	1144	60.9	863.9	46.0

SEIFA, Socio-Economic Indexes for Areas.

and 95% CIs. We considered results significant where p<0.05 and commented on all results where p<0.10.

RESULTS

n=1983 responses were analysed. Weighted and unweighted sample demographics are shown in table 3. Weights primarily affected distributions in self-reported health, chronic health condition or disability status and speaking languages other than English at home.

Support for development of AI

Logistic regression results are displayed in figure 1 with weighted proportions in online supplemental file 2. Overall, 56.7% of the weighted sample supported the development of AI. Support was significantly higher

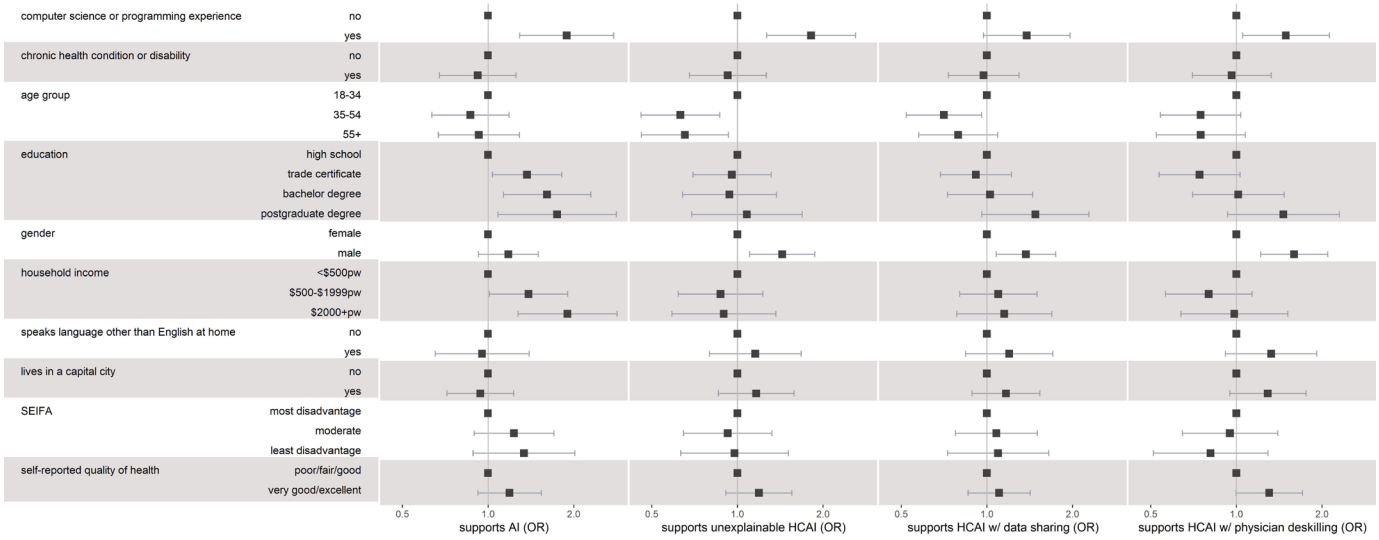

Figure 1 OR plot of weighted logistic regression results. Error bar indicates 95% CI. Index categories displayed with OR=1. Plots indicate (1) participants' level of support for artificial intelligence (AI), (2) participants' level of support for unexplainable AI in healthcare, (3) participants' support for AI in healthcare that necessitates sharing data and (4) participants' support for healthcare artificial intelligence (HCAI) that leads to physician deskilling. pw, per week; SEIFA, Socio-Economic Indexes for Areas.

among those with computer science experience (weighted proportion supportive=72.1%; OR=1.89; p=0.001) compared with those without such experience; those with moderate (55.6%; OR=1.39; p=0.043) or high (66.3%; OR=1.90; p=0.002) household incomes compared with those with low income; and those with trade certificates/diplomas (57.4%; OR=1.37; p=0.028), bachelor's degrees (65.6%; OR=1.61; p=0.008) and postgraduate degrees (69.0%; OR=1.75; p=0.022) compared with those with only high school-level education.

Support for the development of HCAI and trade-offs

Participants were asked to consider whether they supported the development of HCAI in three scenarios. Across the weighted sample, only 27.0% were supportive of HCAI that led to physician deskilling, 28.7% were supportive of unexplainable HCAI and 41.9% were supportive of HCAI that necessitated sharing personal data. Logistic regression results are displayed in figure 1.

Support for unexplainable HCAI was significantly higher among those with computer science experience (43.4%; OR=1.82; p=0.001) and males (32.5%; OR=1.44; p=0.007). Support was significantly lower among those aged 35–54 (25.3%; OR=0.63; p=0.005) and those aged 55+ (25.0%; OR=0.65; p=0.018) compared with those aged 18–34 (36.4%).

Support for AI that necessitates data sharing was significantly higher among males (46.0%; OR=1.37; p=0.011). Participants aged 35–54 (38.4%; OR=0.71; p=0.025) were less likely than those aged 18–34 (48.4%) to be supportive of HCAI that necessitates data sharing.

Support for HCAI that leads to physician deskilling was significantly higher among those with computer science experience (40.0%; OR=1.49; p=0.025) and males (31.6%; OR=1.60; p=0.001).

The analysis did not show an association between household income, living in areas with less social disadvantage, living in a capital city, speaking languages other than English at home or having a chronic health condition/disability and support for the HCAI trade-offs.

Importance of different features in AI-integrated healthcare

Participants were asked to respond to a series of questions about the importance of various aspects of HCAI implementation. Logistic regression results can be found in figure 2 and weighted proportions for each subgroup can be found in online supplemental file 3. Across all sociodemographic groups, accuracy was the feature most regarded as important, and reducing costs to the healthcare system was least likely to be regarded as important followed by speed.

Socioeconomic characteristics

Socioeconomic factors had a little effect on perceived importance of the features. Having a high (>$2000 per week) income had a weak positive effect on perceived importance of reducing costs to the healthcare system (64.5%; OR=1.44; p=0.073). SEIFA was not associated with perceived importance for any of the features.

Demographic characteristics

Demographic characteristics had some associations with perceived importance of the features. Those who spoke languages other than English at home were significantly less likely to regard explainability (68.0%; OR=0.66; p=0.035) and equity (65.1%; OR=0.66; p=0.035) as very/extremely important. They were also perhaps less likely to perceive accuracy (77.7%; OR=0.65; p=0.056) and accountability (70.6%; OR=0.70; p=0.074) as very/extremely important. Those aged over 55 were more

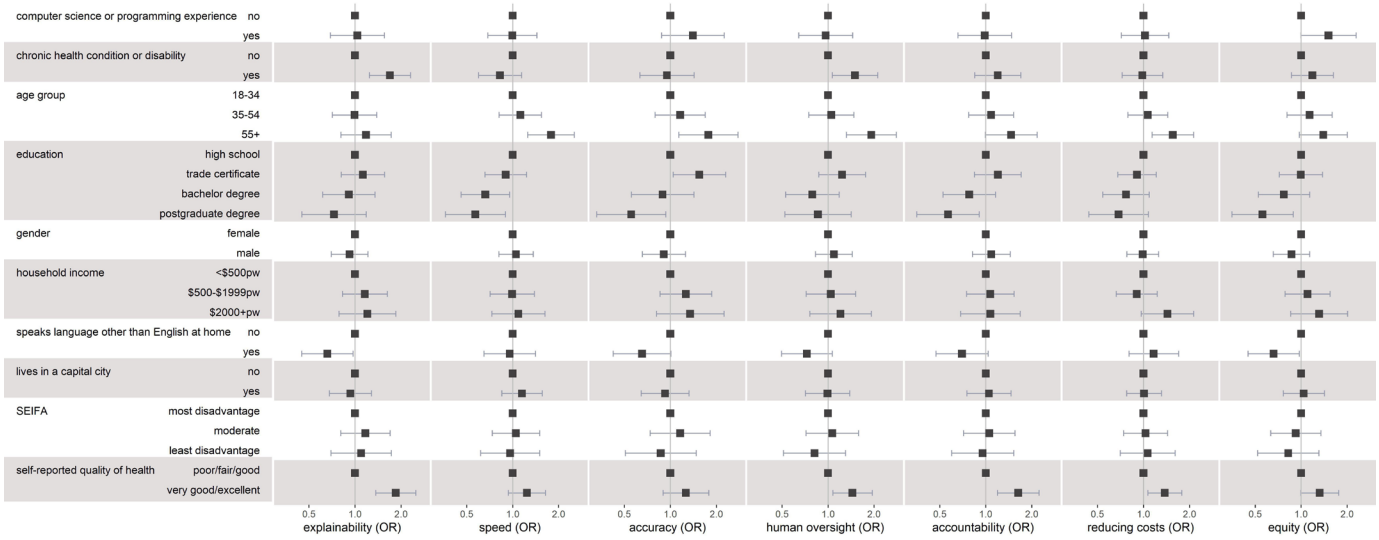

Figure 2 OR plots of weighted logistic regression results. Error bar indicates 95% CI. Index categories displayed with OR=1. Plots indicate level of importance attributed to each aspect of artificial intelligence (AI)-enabled care. pw, per week; SEIFA, Socio-Economic Indexes for Areas.

likely than those aged 18–34 to perceive all features as very important, particularly human oversight (85.0%; OR=1.92; p=0.001); however, this effect was not significant for equity and explainability. Gender and living in a capital city had no significant association with any of the features.

Educational characteristics

Those with postgraduate degrees were less likely than those with a high school-level education to see accuracy (73.9%; OR=0.55; p=0.027), equity (64.0%; OR=0.56; p=0.014), speed (61.1%; OR=0.57; p=0.015) and accountability (61.1%; OR=0.57; p=0.018) as very/extremely important. Those with computer science or programming experience were slightly more likely to see equity (76.0%; OR=1.51; p=0.052) as very/extremely important.

Health-related characteristics

Those with higher self-reported health were significantly more likely to perceive all features as important, except for equity (at p=0.056), speed and accuracy. Those who identified as having a chronic health condition were significantly more likely than those who did not to perceive explainability (81.1%; OR=1.69; p=0.001) and human oversight (83.2%; OR=1.5; p=0.02) as very/extremely important.

DISCUSSION

In this study we examined sociodemographic differences in preference for healthcare AI using a large weighted Australian sample that was calibrated to the LIA probability sample using a range of behavioural and lifestyle questions, as well as major sociodemographic variables. Overall, 56.7% (95% CI 53.8%–59.0%) of the participants were supportive of the development of AI, slightly lower than results from another recent Australian study that also used an online panel, which found 62.4% were supportive.[15] In a separate analysis of the same AVA-AI survey, combining the LIA probability sample results with the online panel results,[14] it was found that 60.3% (95% CI 58.4%–62.0%) of Australians were supportive of the development of AI. In the unweighted non-probability sample, 54.8% (95% CI 52.5%–57%) of participants supported the development of AI, suggesting that the use of an extensive set of variables in the weighting led to some improvement, but the potential of self-selection in online panels may not have been corrected fully by the sophisticated weighting methodology.

Similar to Zhang and Dafoe's[10] study in the USA, we found that support for the development of AI was higher among those with computer science experience, higher levels of education and higher household incomes. It has been suggested that support for AI is lower among groups with less education and more social disadvantage, whose livelihoods may be more threatened by automation.[10 12] The potential for AI to threaten people's livelihoods through taking jobs appears to be a poignant concern in Australia, where Selwyn *et al*[15] found that the prospect of automation and job loss was the most commonly mentioned fear among their Australian sample. Results from our survey appear to support these findings, where metrics for social advantage (ie, household income and education) were strongly associated with support for development of AI.

The sociodemographic characteristics associated with support for HCAI were different from those associated with support for AI in general. The items assessing support for HCAI required participants to consider whether they supported the development of HCAI, on balance, when it involved a trade-off (lack of explainability, data sharing or physician deskilling). For each of the HCAI questions, household income and education

were no longer predictors of support. For example, 66.3% of the weighted sample with incomes >$2000 per week supported the development of AI in general, and only 30.5% supported the development of unexplainable HCAI. In contrast, 45.9% of those with incomes <$500 per week supported AI in general and 29.7% supported the development of unexplainable HCAI. This suggests that measures of socioeconomic advantage are linked to a general support of the development of AI, but when assessing specific and potentially harmful applications of HCAI, there is a low level of support regardless of socioeconomic characteristics.

Qualitative research on HCAI with members of the public has found that attitudes towards HCAI are shaped by complex evaluations of the alignment of the technologies with the values of medicine.[21] If this is the case, then support for HCAI may be driven less by economic values and more by values relating to healthcare.

The characteristics that we found to be consistent predictors of support for HCAI and their specified trade-offs were having computer science experience, being male and being aged 18–34. Similarly, Zhang and Dafoe[10] found that younger people and those with computer science degrees expressed less concern about AI governance challenges than those who were older or did not have computer science qualifications.

Being male, having computer science experience and being in a younger age category were three characteristics among those Selwyn *et al*[15] found were associated with higher levels of familiarity with AI. It is possible that subgroups more familiar with AI are perhaps more tolerant of its risks. However, the Selwyn and colleagues' study did not control for potential confounding relationships between age, gender and computer science experience so it is unclear from this work whether age and gender were indeed associated with greater familiarity with AI or whether a greater proportion of their younger male sample also had computer science experience, which may be more likely associated with higher levels of familiarity with AI. The relationship between familiarity with AI and tolerance of its risks may warrant further investigation.

Our investigation into subgroup differences in the perceived importance of features of HCAI found that accuracy was regarded as particularly important by all subgroups. This differs from Ploug *et al*[22] who found, in a choice experiment in Denmark, that factors like explainability, equity and physicians being responsible for decisions were regarded as more important than accuracy. The Danish experiment, however, offered the qualifier that the algorithm would at least be as accurate as a human doctor, whereas our questionnaire did not. Further research could test whether algorithmic performance is more important than other features in circumstances where there are no assurances that the algorithm is as accurate as a human doctor.

Health-related characteristics such as self-reported health and having a chronic health condition or disability had a strong effect on perceived importance attributed to traditionally human aspects of healthcare like explainability, human oversight and accountability. This result is echoed by Richardson *et al*'s[21] finding that people's discussions about the value of HCAI were often framed by their previous experiences with the healthcare system. Participants with complex health needs may have been more inclined to reflect on whether automated systems could meet all aspects of those needs.

Subgroups that were more likely to be supportive of HCAI were not necessarily more likely to see the features of care that they were trading off as less important. While those who identified as male, those aged 18–34 and those with computer science or programming experience were more likely to support the development of unexplainable AI in healthcare, they were just as likely as others to perceive explainability ('*knowing why a decision is made*') as an important aspect of AI-integrated care. This hints at a complex relationship between people's support for the development of HCAI and their willingness to make compromises to their healthcare.

Limitations

Given the quickly shifting landscape around AI, it is possible that public support for AI has changed in the 2 years since the questionnaire was administered. In addition, the AVA-AI survey includes an online panel obtained by non-probability sampling, which is subject to self-selection biases. The weighting methodology assists in reducing these effects by accounting for more than basic demographic variables, such as age by education, gender, household structure, language spoken at home, self-reported health, early adopter status and television streaming. Any selection effects due to the prediction variables included in the analysis are also accounted for. However, it is possible that support for HCAI is mediated by confounding factors not considered in the weighting methodology or included in the analysis.

One key population that were not represented in the study were those who identified as a gender outside of the male/female binary. Only one participant identified as a gender outside of the binary and was excluded from the analysis due to insufficient participant numbers to form a third gender category. Given that support for AI is lower among certain marginalised groups, consulting gender diverse individuals about their support for AI is an important consideration for future research.

Finally, the present study is a cross-sectional analysis which cannot infer causation between any of the predictor and outcome variables. While we found an association between certain sociodemographic characteristics such as education, and outcomes such as level of support for AI, we cannot ascertain the reasons for this association. These reasons are likely complex and multifaceted and should be explored in further research.

CONCLUSION

Respondents who reported having greater ill health or disability were more likely to consider human aspects of healthcare, such as explainability, human oversight and accountability, as important. While factors indicating socioeconomic advantage (higher income, higher education) were associated with general support for AI, these factors were not necessarily related to support for HCAI scenarios. Instead, support for HCAI scenarios was higher among males, younger people and those with computer science or programming experience. Based on other research, these groups may have a higher level of familiarity with AI. Further research should examine the relationship between familiarity with AI and support for the development of AI.

Author affiliations
[1]Australian Centre for Health Engagement, Evidence and Values, School of Health and Society, Faculty of the Arts, Social Sciences and Humanities, University of Wollongong, Wollongong, New South Wales, Australia
[2]School of Mathematics and Applied Statistics, Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, New South Wales, Australia
[3]Centre for Sample Survey Methodology, National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, New South Wales, Australia

Twitter Emma Kellie Frost @EmmaKFrost and Yves Saint James Aquino @yvessj_aquino

Acknowledgements Statistical consulting was provided by Dr Brad Wakefield at the National Institute for Applied Statistics Research Australia (NIASRA) Statistical Consulting Centre.

Contributors EKF contributed to the design of the study, cleaned the data, conducted the analysis, drafted the manuscript, and acts as guarantor. PO'S contributed to the design of the survey instrument, contributed to the design of the study, oversaw the analysis and edited the manuscript. DS contributed to the design of the survey and edited the manuscript. AB-M and YSJA edited the manuscript. SMC contributed to the design of the survey instrument, contributed to the design of the study and edited the manuscript. All authors approved the final version of the manuscript for publication.

Funding Funding for this survey was provided by the University of Wollongong's Global Challenges Program.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by University of Wollongong HREC (2019/458). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD
Emma Kellie Frost http://orcid.org/0000-0002-5893-1399

REFERENCES

1 Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Illustrated edition*. New York: Basic Books, 2019.
2 Hajkowicz S, Karimi S, Wark T, *et al*. *Artificial Intelligence: Solving problems, growing the economy and improving our quality of life, Data61*. CSIRO, 2019.
3 Attia ZI, Noseworthy PA, Lopez-Jimenez F, *et al*. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019;394:861–7.
4 Esteva A, Topol E. Can skin cancer diagnosis be transformed by AI *Lancet* 2019;394:1795.
5 Kim JOR, Jeong Y-S, Kim JH, *et al*. Machine learning-based cardiovascular disease prediction model: a cohort study on the Korean national health insurance service health screening database. *Diagnostics (Basel)* 2021;11:943.
6 Carter SM, Rogers W, Win KT, *et al*. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast* 2020;49:25–32.
7 Kurani S, Inselman J, Rushlow DR, *et al*. Abstract 9213: effect of ai-enabled ECG screening for low EF in patients with different socioeconomic status. *Circulation* 2021;144:A9213.
8 Seyyed-Kalantari L, Zhang H, McDermott MBA, *et al*. Underdiagnosis bias of artificial intelligence Algorithms applied to chest Radiographs in under-served patient populations. *Nat Med* 2021;27:2176–82.
9 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
10 Zhang B, Dafoe A. 2 general attitudes toward AI | Artificial intelligence: American attitudes and trends. 2019. Available: https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/general-attitudes-toward-ai.html [Accessed 25 Jun 2021].
11 Yakar D, Ongena YP, Kwee TC, *et al*. Do people favor artificial intelligence over physicians? A survey among the general population and their view on artificial intelligence in medicine. *Value Health* 2022;25:374–81.
12 Kovarik CL. Patient perspectives on the use of artificial intelligence. *JAMA Dermatol* 2020;156:493–4.
13 Social Research Centre. Life in Australia study. Available: https://www.srcentre.com.au/our-research/life-in-australia-study [Accessed 21 Apr 2021].
14 Isbanner S, O'Shaughnessy P, Steel D, *et al*. The adoption of artificial intelligence in health care and social services in Australia: findings from a Methodologically innovative national survey of values and attitudes (the AVA-AI study). *J Med Internet Res* 2022;24:e37611.
15 Selwyn N, Cordoba BG, Andrejevic M, *et al*. *AI for social good - Australian public attitudes toward AI and society*. 2020: 76.
16 Australian Bureau of Statistics. Main features - Socio-Economic Indexes for Areas (SEIFA) 2016. 2018. Available: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2033.0.55.001~2016~Main%20Features~SOCIO-ECONOMIC%20INDEXES%20FOR%20AREAS%20(SEIFA)%202016~1 [Accessed 06 Sep 2022].
17 Cameron JJ, Stinson DA. Gender (Mis)Measurement: guidelines for respecting gender diversity in psychological research. *Soc Personal Psychol Compass* 2019;13.
18 Barnier J, Briatte F, Larmarange J. questionr: Functions to make surveys processing easier. 2022. Available: https://CRAN.R-project.org/package=questionr [Accessed 08 Sep 2022].
19 R Core Team. R: A language and environment for statistical computing. Vienna, Austria R Foundation for Statistical Computing; 2013. Available: http://www R-project org/
20 Lumley T. Survey: analysis of complex survey samples. R package version 3.35-1, 2019; 2020.
21 Richardson JP, Curtis S, Smith C, *et al*. A framework for examining patient attitudes regarding applications of artificial intelligence in Healthcare. *Digit Health* 2022;8:20552076221089084.
22 Ploug T, Sundby A, Moeslund TB, *et al*. Population preferences for performance and explainability of artificial intelligence in health care: choice-based Conjoint survey. *J Med Internet Res* 2021;23:e26611.

BMJ Health & Care Informatics

Anticipating artificial intelligence in mammography screening: views of Swedish breast radiologists

Charlotte Högberg ,[1] Stefan Larsson,[1] Kristina Lång[2,3]

To cite: Högberg C, Larsson S, Lång K. Anticipating artificial intelligence in mammography screening: views of Swedish breast radiologists. *BMJ Health Care Inform* 2023;**30**:e100712. doi:10.1136/bmjhci-2022-100712

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100712).

Received 05 December 2022
Accepted 08 May 2023

[1]Department of Technology and Society, Lund University Faculty of Engineering, Lund, Sweden
[2]Department of Translational Medicine, Diagnostic Radiology, Lund University Faculty of Medicine, Lund, Sweden
[3]Unilabs Mammography Unit, Skåne University Hospital Lund, Malmö, Sweden

Correspondence to
Charlotte Högberg;
charlotte.hogberg@lth.lu.se

ABSTRACT

Objectives Artificial intelligence (AI) is increasingly tested and integrated into breast cancer screening. Still, there are unresolved issues regarding its possible ethical, social and legal impacts. Furthermore, the perspectives of different actors are lacking. This study investigates the views of breast radiologists on AI-supported mammography screening, with a focus on attitudes, perceived benefits and risks, accountability of AI use, and potential impact on the profession.

Methods We conducted an online survey of Swedish breast radiologists. As early adopter of breast cancer screening, and digital technologies, Sweden is a particularly interesting case to study. The survey had different themes, including: attitudes and responsibilities pertaining to AI, and AI's impact on the profession. Responses were analysed using descriptive statistics and correlation analyses. Free texts and comments were analysed using an inductive approach.

Results Overall, respondents (47/105, response rate 44.8%) were highly experienced in breast imaging and had a mixed knowledge of AI. A majority (n=38, 80.8%) were positive/somewhat positive towards integrating AI in mammography screening. Still, many considered there to be potential risks to a high/somewhat high degree (n=16, 34.1%) or were uncertain (n=16, 34.0%). Several important uncertainties were identified, such as defining liable actor(s) when AI is integrated into medical decision-making.

Conclusions Swedish breast radiologists are largely positive towards integrating AI in mammography screening, but there are significant uncertainties that need to be addressed, especially regarding risks and responsibilities. The results stress the importance of understanding actor-specific and context-specific challenges to responsible implementation of AI in healthcare.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Radiologists believe that artificial intelligence (AI) will have a major impact in their field, and clinical retrospective studies of AI in mammography screening show promising results.

WHAT THIS STUDY ADDS

⇒ The social, ethical and legal aspects of integrating AI in mammography screening are underexplored, and by investigating the views of breast radiologists, this study provides important insights for a responsible approach to AI in mammography screening.

⇒ The study shows that most Swedish breast radiologists are positive about integrating AI in mammography screening, especially those with a heavy screen-reading workload. However, there is no unified vision of how AI should be used in the screening-work flow, and there is high uncertainty, and diverse views, on important aspects such as potential risks involved, and which actor(s) are liable for medical decision-making, particularly when AI is used as stand-alone reader.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study adds to the emerging body of research on AI in medical decision-making, taking into account contextual and actor-specific factors, and emphasises the social, ethical and legal unclarities of integrating AI into mammography screening, that must be addressed.

INTRODUCTION

In radiology, the use of artificial intelligence (AI) is rapidly evolving. One area targeted as especially promising is mammography screening.[1–3] The benefit of population-based screening is early breast cancer detection, reducing mortality and morbidity. This is a benefit balanced by the harm of false positives, overdiagnosis and false negatives.[4 5] The vast majority of individuals who are screened do not have breast cancer, however, screen examinations are, in European guidelines, recommended to be double-read to ensure a high sensitivity.[6] The hope is that integrating AI will result in a more efficient screening with reduced workload and a potentially higher accuracy. By adapting single-reading and double-reading to AI risk scores, or combining it with automated reading of low-risk examinations, it is suggested that the workload may be reduced by up to 63%.[7] In theory, reducing the number of exams that are double-read will lead to fewer false positives.[8] Retrospective studies have also shown that AI could potentially lower the

false-negative rate,[9 10] but prospective studies are still needed to understand the real impact of AI.[11]

Beyond clinical aspects, there are unresolved issues regarding the ethical, social and legal consequences of integrating AI into healthcare.[12 13] These include how to safeguard values of medical ethics such as fairness, accountability and transparency.[14 15] These matters are perceived as some of the greatest hurdles of implementing AI in radiology.[16–19] In response, standards for AI in radiology are stressed, including the equal distribution of benefits and harms between stakeholders, transparency of AI-systems, curtailing bias in decision-making, and that accountability should remain with humans.[12]

The expectation that AI will change the field of radiology in the near future is highly prevalent among radiologists, trainees and medical students.[16–18] Still, different stakeholders' notions of the challenges are in need of more exploration.[20–22] Prior studies show high willingness to use AI in clinical practice, but this could differ depending on subfield. While included in studies as one subdiscipline of many, not much focus has been dedicated specifically to breast radiologists, a group likely to be involved in AI-implementation on a large scale, and with experience of the particular conditions of the screening process.[16 23] In addition, mammography screening is a major medical intervention that affects a large part of the population, and social and ethical implications of integrating AI in this context are underexplored.[19 22] Therefore, we are examining the views of breast radiologists. Moreover, Sweden is an especially relevant case, as it is one of the most digitalised countries in the European Union,[24] as well an early adopter of population-based breast cancer screening and with ongoing pioneering prospective trials on AI in screening.[25]

This study investigated Swedish breast radiologists' views on the use of AI in mammography screening and their perceptions of the risks, benefits and responsibilities of actors involved, and its impact on the profession.

METHOD

An online survey (using Sunet Survey) was distributed to the Swedish Society of Breast Imaging (SSBI), in which the vast majority of Swedish breast radiologists are members. The survey was conducted over the course of 1 month in the late fall of 2021. Informed consent was obtained before answering the survey, by click-response. The questionnaire contained 45 questions falling under different themes. Besides background questions used to establish respondent characteristics, questions were chosen due to their relevance for the social, ethical and legal issues of AI implementation. This included (but was not limited to): attitudes about AI-supported mammography screening, responsibility of AI-use and the future professional impact of AI integration (see online supplemental appendix A). Background questions had categorical response options. Two questions only had free-text response option. The remaining questions had Likert-scale response options, representing degrees (to a low degree, to a somewhat low degree, uncertain, to a somewhat high degree, to a high degree) or attitudes (negative, somewhat negative, uncertain, somewhat positive, positive). In addition, the respondents had the opportunity to provide free-text comments.

Results were analysed using descriptive statistics and are presented in percentages and frequencies. Correlation analyses were performed by Spearman's r, with 95% CI and p values of <0.05 considered statistically significant. In addition, cross-tabulations were used to cover more correlations. Statistical analyses were performed using IBM SPSS Statistics for Mac (V.28.0, IBM). All free-text responses and comments were analysed using an inductive approach and used as method triangulation complementing the quantitative results. Since comments and the two free-text questions were optional, not all respondents' views are accounted for; still, they provide valuable means for obtaining a deeper understanding.

RESULTS

Out of the 105 members of the SSBI, 47 answered the survey (response rate: 44.8%). Of these, 25 were females (53.2%), and the majority of the respondents were older (66.5%>50 years of age), most had long experience in breast imaging (70.2%>11 years of experience) and a high reading-volume was fairly common (38.3% performed >10 000 screen-readings per year). A majority (n=33, 73.3%) of the respondents reported that they sometimes, often or always had difficulties finding time to do screen-readings. More respondents estimated to have higher literacy of technology in everyday life and at work in general, than of AI. Most (n=18, 38.3%) estimated their AI literacy to be neither high nor low, and 25.5% that it was somewhat high or high. Correspondingly, 21.3% had extensive or somewhat extensive experience of using AI in their work, while nearly half (n=22, 46.8%) had no experience (table 1).

Attitudes, benefits and risks of AI in mammography screening

Positive views and potential benefits

The breast radiologists were, to a large extent, positive towards AI-supported mammography screening; 80.8% (n=38) being somewhat positive or positive (table 2, figure 1A). Comments suggest that AI is perceived as a good complement and solution to the scarcity of breast radiologists. Furthermore, having difficulties finding the time to perform screen-readings correlated with a positive attitude towards AI-supported screening (Spearman's r=0.367, 95% CI, p=0.013, figure 2). A correlation between self-estimated literacy of AI and attitude could not be established (p=0.825).

A majority (n=37, 78.7%) of the respondents believed that there were potential benefits in using AI-supported screening, to a somewhat high or high degree (figure 1B). Benefits specified in the comments were improved detection and consistency in screen-reading. The respondents

Table 1 Background characteristics of participating breast radiologists

Characteristics of respondents	N (%)
Age (Q1)	(N=47)
<30	0 (0)
31–40	3 (6.4)
41–50	13 (27.7)
51–60	13 (27.7)
>60 years	18 (38.3)
Gender (Q2)	(N=47)
Female	25 (53.2)
Male	22 (46.8)
Experience of breast radiology (Q3)	(N=47)
<5 years	5 (10.6)
5–10 years	9 (19.1)
11–20 years	10 (21.3)
21–30 years	11 (23.4)
>30 years	12 (25.5)
Screen readings approx. performed per year (Q4)	(N=47)
None	2 (4.3)
<2000	4 (8.5)
2000–5 000 000	5 (10.6)
5000–10 000 000	18 (38.3)
>10 000	18 (38.3)
Difficulties finding time to perform screen-readings (Q5)	(N=45)
Never	4 (8.9)
Seldom	8 (17.8)
Sometimes	19 (42.2)
Often	8 (17.8)
Always	6 (13.3)
Self-estimated technology literacy, everyday life (Q6)	(N=47)
Low	0 (0)
Somewhat low	1 (2.1)
Neither high nor low	20 (42.6)
Somewhat high	19 (40.4)
High	7 (14.9)
Self-estimated technology literacy, work (Q7)	(N=47)
Low	0 (0)
Somewhat low	1 (2.1)
Neither high nor low	17 (36.2)
Somewhat high	21 (44.7)
High	8 (17.0)
Self-estimated AI literacy (Q8)	(N=47)
Low	4 (8.5)
Somewhat low	13 (27.7)
Neither high nor low	18 (38.3)
Somewhat high	8 (17.0)
High	4 (8.5)
Experience of using AI in breast radiology (Q12)	(N=47)
None	22 (46.8)
Little	9 (19.1)
Somewhat little	6 (12.8)
Somewhat large	6 (12.8)
Large	4 (8.5)

AI, artificial intelligence.

favoured using AI as a replacement of 1 reader in double-reading (n=21, 44.7%) or in addition to 2 human readers (n=14, 29.8%) (table 2). A wish to combine triage, reader replacement and detection support were also mentioned in the comments.

Negative views and potential risks

Nearly one-fifth of respondents (n=9, 19.2%) were negative/somewhat negative or uncertain about AI-supported screening (table 2). In the comments, negative attitudes refer to experiencing AI as linked to large numbers of false positives (due to a high sensitivity for calcifications), difficulty in interpreting AI-assessments and the risk of an increased workload, as expressed by one respondent:

> *It was annoying to have to go back and assess different AI findings of benign things all the time […] that I would normally not have had to put any energy into assessing. It made the work slower and disrupted the work pace, leading to more exhaustion* [P12].

The views concerning potential risks of integrating AI were diverse (figure 1B). Comments revealed that, besides medical risks, some feared AI would lead to a deterioration of working conditions, an increase in false positives and interpretation load, and loss of competence due to a lack of continuous training on healthy mammograms:

> *Consultation hours with ultrasounds and biopsies are often heavily booked with worried patients. Working whole days like that would be hard.* [P12]

> *There is a risk that AI detects findings that are obviously benign, which will take time and effort to investigate and prove. Some changes that are unquestionable to a radiologist, AI can't see* [P23].

Other comments stressed legal and ethical risks. One case mentioned, was if a physician disregards an AI finding that later turns out to be a cancer. The respondent suggested radiologists will be put in that situation 'all the time' since AI detects so many findings.

Table 2 General attitudes and perceived potential benefits and risks of AI-supported mammography screening

Attitudes, perceived benefits and risks	N (%)
Attitude towards AI-supported mammography screening (Q9)	(N=47)
Positive	11 (23.4)
Somewhat positive	27 (57.4)
Uncertain	6 (12.8)
Somewhat negative	1 (2.1)
Negative	2 (4.3)
Preferred use of AI in mammography screening (Q13)	(N=47)
AI as triage tool	6 (12.8)
AI as stand-alone reader	2 (4.3)
AI as replacement of one in double reading	21 (44.7)
AI as addition to double reading	14 (29.8)
Not at all	4 (8.5)
Potential benefits of AI-supported screening (Q10)	(N=47)
To a high degree	13 (27.7)
To a somewhat high degree	24 (51.1)
Uncertain	6 (12.8)
To a somewhat low degree	2 (4.3)
To a low degree	2 (4.3)
Potential risks of AI-supported screening (Q11)	(N=47)
To a high degree	6 (12.8)
To a somewhat high degree	10 (21.3)
Uncertain	16 (34.0)
To a somewhat low degree	14 (29.8)
To a low degree	1 (2.1)
Perceived risk of overconfidence in AI assessments (Q15)	(N=47)
To a high degree	4 (8.5)
To a somewhat high degree	9 (19.1)
Uncertain	20 (42.6)
To a somewhat low degree	13 (27.7)
To a low degree	1 (2.1)
Perceived risk of non-representative training data (Q38)	(N=47)
To a high degree	2 (4.3)
To a somewhat high degree	9 (19.1)
Uncertain	24 (51.1)
To a somewhat low degree	9 (19.1)
To a low degree	3 (6.4)
Perceived risk of inferior AI performance on certain risk groups or specific type of cases (Q39)	(N=47)
To a high degree	6 (12.8)
To a somewhat high degree	13 (27.7)
Uncertain	22 (46.8)
To a somewhat low degree	5 (10.6)
To a low degree	1 (2.1)

AI, artificial intelligence.

About half of the respondents (n=24, 51.1%) were uncertain as to whether there are risks in AI-models being trained on data that are not representative of the population to which they are applied. Many were also uncertain as to whether AI-models perform poorly on risk groups or certain types of cases (n=22, 46.8%) (table 2). Cases perceived as possibly more difficult for AI to assess included; dense breasts, atypical soft tissue masses without calcification, architectural distortion, developing asymmetric density, postoperative changes or young individuals with hereditary risk.

Accountability of AI-use

When AI is used in addition to radiologists in screen reading, most the respondents (n=31, 65.9%) considered the radiologist to be responsible for the assessments to a high/somewhat high degree, but 21.3% (n=10) were uncertain (figure 3A). When AI is used as a stand-alone reader, the radiologist (eg, in terms of oversight) was considered responsible to a high/somewhat high degree only by 12.8% (n=6) (figure 3B). The healthcare provider was, to a larger extent, considered responsible when AI is stand-alone reader, compared with when it is used in addition to radiologist(s). This was also the case regarding the responsibility of developers of AI-systems (figure 3).

To answer whether agency was ascribed to the AI-system, as is common in everyday discussions about AI, we included it as an option among liable actors. When used in addition to radiologist(s), 38.3% (n=18) of the respondents considered the AI-system to be responsible to a high/somewhat high degree. When used as stand-alone reader, the number was larger (n=23, 48.9%) and about one-third of the respondents were uncertain (n=14, 29.8%) (figure 3). Perceived shared responsibility was less prevalent when AI is used as a stand-alone reader. Uncertainty and urgency on the issue of responsibility emerged in the comments: *This is the most difficult part. Who takes responsibility? Healthcare should do it, probably, but it is actually the AI-system and the AI-developer who should be accountable for the result* [P24].

Impact on the profession

Nearly half of the breast radiologists in the sample (n=21, 44.7%) believed that integrating AI in mammography screening would encompass substantial differences in comparison to other previously introduced technologies (such as digital mammography and tomosynthesis), to a high/somewhat high degree. A comment suggested the reason for this was that previous technologies aimed to

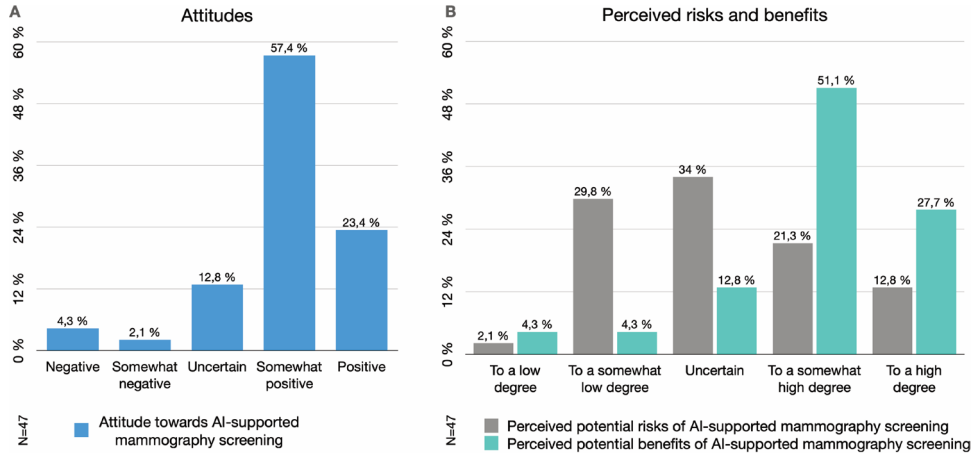

Figure 1 Attitudes. (A) The Swedish breast radiologists' attitude towards AI-supported mammography screening (Q9). (B) The perceived degree of benefits and risks of AI-supported mammography screening (Q10 and Q11). AI, artificial intelligence.

improve image quality, while AI is about delegating assessments to the technology. However, more than one-third of the respondents (n=17, 36.2%) were uncertain as to whether there were any substantial differences of introducing AI.

Moreover, there was a mix of viewpoints regarding how integrating AI might impact the role of the breast radiologist. Most commonly (n=20, 42.6%) AI was believed to have no impact, while nearly a third of respondents (n=15, 31.9%) thought it would strengthen/somewhat strengthen the role of breast radiologists and 25.5% (n=12) that it would weaken/somewhat weaken it. Only 21.3% (n=10) believed the use of AI would make it easier to recruit new breast radiologists to a high/somewhat high degree.

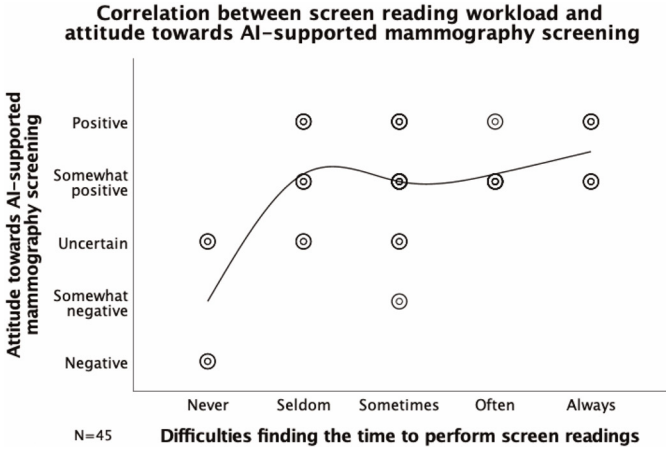

Figure 2 Correlation between the Swedish breast radiologists' attitude towards AI supported mammography screening (Q9) and their difficulties to find the time to perform screen-readings (Q5) (Spearman's r=0.367, 95% CI, p=0.013). The results suggest that the Swedish breast radiologists' experienced screen-reading workload correlates with their attitude towards AI-supported mammography screening. AI, artificial intelligence.

Relation to screening participants

The question about whether implementing AI-supported mammography screening would impact the relationship with screening participants was answered using free-text responses. Out of the total sample, 32 respondents answered, and both positive and negative outlooks were articulated. Some stated that the use of AI would increase the participants' trust, and improve working conditions and thereby also the relationship with caretakers. Other respondents suggested that trust would decrease and introducing AI would 'lead to chaos' and 'waste everyone's time'. Several highlighted the importance of AI systems being valid and trustworthy, and to be able to convey that trustworthiness to relevant actors. Some respondents also emphasised the significance of having radiologists in charge of AI implementation, management and quality control.

Technological development

How the profession will evolve, in light of current technological development, provided a mix of viewpoints. Some pointed to socioeconomic factors, such as: *[I] think that AI will be implemented in screening considering the economic benefits it could have for the employers* [P33]. Several respondents voiced insecurities and expressed reservations:

> *It probably cannot be avoided in the long run and should be able to provide more time for what needs to be investigated or acted upon. I am a bit worried about the loss of knowledge of the "normal breast as background"* [P2].

Still, others emphasised that they considered AI as not yet reaching an acceptable performance level: *until AI becomes good enough, it will be a long way* [P14]. However, other responses expressed hopes of what AI-integration could bring; easing screen-reading workload, improving diagnostics and healthcare quality, with statements like: *AI will be able to sort out the easier cases, decrease workload, and help to find more cancers* [P36]. Moreover, some comments emphasised AI's supporting qualities:

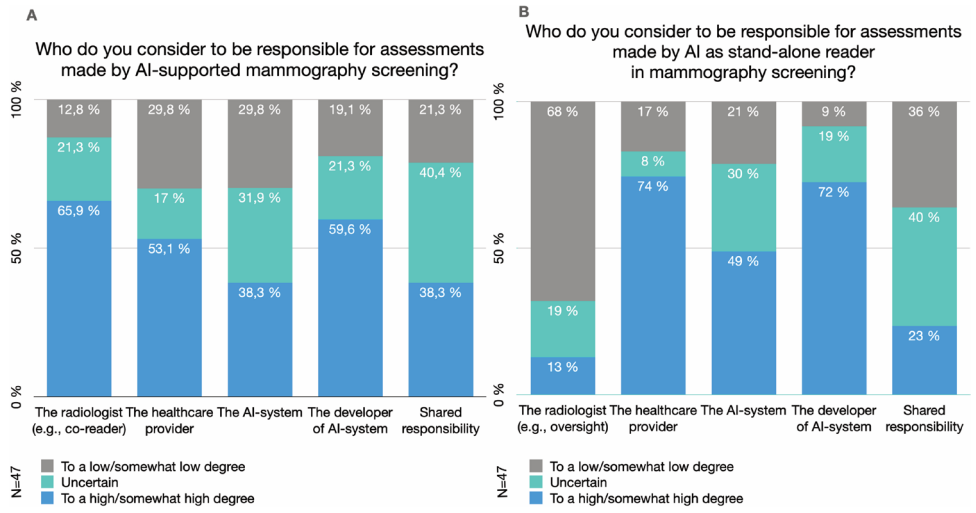

Figure 3 Accountability. (A) The Swedish breast radiologists' perceived levels of accountability of different actors for assessments made by AI-supported mammography screening (Q17–Q21). (B) The Swedish breast radiologists' perceived levels of accountability of different actors for assessments made by AI as stand-alone reader in mammography screening (Q22–Q26). AI, artificial intelligence.

> *It feels good to be able to be supported by AI in the screening, it could be nice to have a second reader (AI) whom never loses concentration. Considering the screening volumes and the scarcity of breast radiologists it feels good that AI can complement us.* [P16]

DISCUSSION

In this study, we have investigated Swedish breast radiologists' views on the use of AI in mammography screening. The respondents were, to a large extent, positive towards the integration of AI in screen reading, especially those having difficulties finding the time to perform screen-reading. This could explain the slightly more positive attitude, compared with general studies on radiologists' attitudes towards AI.[23 26 27] We could not establish a correlation between attitude and AI-literacy, prevalent in previous general studies.[23] However, it needs to be taken into account that our sample represents a relatively small number of individuals. Those more opinionated about AI could also be more inclined to answer the survey, possibly inducing bias in the results. The specific context, of mammography screening and the profession of breast radiologists, in a digitally advanced welfare state, however, showcases the importance of considering technological implementations in relation to organisational and socio-economic structures.

Furthermore, we did identify important reservations, factors associated with high uncertainties, and diverse viewpoints, such as regarding liability of AI use. The question is whether established practices need to be adjusted when medical decisions are increasingly supported by automated technologies or AI-systems. Our results point to a somewhat higher perceived responsibility of radiologists in AI-supported radiological practice, compared with previous studies.[20 27] Furthermore, the results show the complexity of accountability when AI enters radiology, how it is contextual, dependent on how AI is used and which actors are included. This further emerges in the insecurities regarding liability for missed cancers, when AI is used as a co-reader or as stand-alone reader, or when radiologists disregard AI findings. The results indicate a perceived shift of responsibility away from the radiologist as automation increases. Additionally, uncertainties regarding the responsibilities of AI-developers (and AI-systems) suggest a need for clarification.[28]

We could not identify one unified vision of a preferred way to use AI in mammography screening. Previously, AI has been expected to be used as second reader and for optimising workflows.[16–18] While using AI as replacement for one reader in double-reading was the most preferred option in our study, many favoured using it as an addition to double-reading or in a combination of uses, suggesting perceived qualities other than workload reduction. Furthermore, the perceived risk of AI deteriorating working conditions might be due to several reasons. Besides a risk of eroded knowledge of the normal breast, reduced screen reading workload might not improve working conditions. While more time for patient-centred care is portrayed as a positive outcome of AI,[27] some perceived screen reading as a welcomed interruption from emotionally burdensome work, which might be lost due to automation. Working with AI-systems also adds layers of interpretation,[29] which could be exhausting. This seems to be perceived as a medical risk, but also as an ethical burden with legal uncertainties.

Additionally, AI in mammography screening needs to be considered in light of previous innovations. Some aspects are not unique for AI, such as contested expertise.[30] However, radiologists, trainees and medical students strongly expect AI to change the field and impact job opportunities, tasks and relationships with

patients.[16–18 20] Our study shows that breast radiologists believed that AI will impact the profession, both positively and negatively. However, most did not believe it would impact the role of the breast radiologist. Few thought it would improve recruitment, possibly due to the idea of AI negatively affecting the professional reputation.[26] Many considered, or were uncertain whether, implementing AI represents a substantial difference in comparison to previous technologies. While new imaging methods aim to improve cancer visibility, AI differs since it involves medical decision-making. This implies that social, ethical and legal aspects have to be addressed, which in turn depends on how AI is incorporated into the clinical workflow. Greater unclarity about accountability seems to be prevalent regarding AI as a stand-alone technique, which was also the least favoured approach. This suggests that physicians are not willing to renounce their responsibility in medical decision-making. In total, our results echo the need for more research on social, ethical and legal matters of integrating AI into radiology and screening.

Strengths and limitations

The main limitations of the study are the specific conditions of the Swedish setting and the small number of respondents. The response rate was satisfactory, but the target population was limited since there are few Swedish breast radiologists. The study's strengths were that the respondents were highly experienced in breast imaging and that half of the group had experience of using AI in breast imaging.

CONCLUSIONS

Breast radiologists in Sweden were largely positive about integrating AI in mammography screening, especially those with a heavy screen-reading workload, citing reduced workload and increased sensitivity as benefits. Still, we identified several concerns and uncertainties that need to be addressed, foremost regarding potential risks – pertaining to medical outcomes, working conditions and the question of liability in medical decision-making when using AI. Furthermore, there is a lack of consensus on the optimal use of AI in the screening workflow. The results emphasise the need to understand actor-specific and context-specific challenges for responsible implementation of AI.

Twitter Charlotte Högberg @Ch_Hogberg and Stefan Larsson @DigitalSocietyL

Acknowledgements We would like to thank the members of Swedish Society of Breast Imaging, SSBI, for taking the time to respond to the survey and for sharing their views. We would also like to thank the board of SSBI for assisting in distributing the survey. Further, we are grateful to the editors of BMJ Health and Care Informatics, and to the two anonymous reviewers for their constructive and insightful comments.

Contributors CH, SL and KL, contributed to the initial idea of the study, the study design, planning and questionnaire. The survey was set up and conducted by CH. Statistical analysis of data was performed by CH.CH, SL and KL, contributed to further analysis and interpretation of data. CH is guarantor.

Funding The study was funded by The Swedish Cancer Society (211631 Pj), The Confederation of Regional Cancer Centers in Sweden (SKR dnr 21/00060), The Swedish Research Council (ref. 2019-00198) and the Swedish Governmental Funding for Clinical Research (ALF) (2020- Projekt0079).

Competing interests KL is on an advisory board for Siemens Healthineers.

Patient consent for publication Not applicable.

Ethics approval This study did not need an ethics approval, in accordance with local regulations, as no sensitive personal information about study participants (the radiologists) was collected. Informed consent was obtained prior to participation.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Data cannot be openly shared, due to respondent anonymity.

ORCID iD

Charlotte Högberg http://orcid.org/0000-0002-1462-6325

REFERENCES

1 Dembrower K, Wåhlin E, Liu Y, *et al*. Effect of artificial intelligence-based Triaging of breast cancer screening mammograms on cancer detection and Radiologist workload: A retrospective simulation study. *Lancet Digit Health* 2020;2:S2589-7500(20)30185-0:e468–74.:.
2 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, *et al*. Stand-Alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI* 2019;111:916–22.
3 McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89–94.
4 Houssami N, Irwig L, Ciatto S. Radiological surveillance of interval breast cancers in screening programmes. *Lancet Oncol* 2006;7:259–65.
5 Loving VA, Aminololama-Shakeri S, Leung JWT. Anxiety and its association with screening mammography. *Journal of Breast Imaging* 2021;3:266–72.
6 Perry N, Broeders M, de Wolf C, *et al*. European guidelines for quality assurance in breast cancer screening and diagnosis. fourth edition -- summary document. *Ann Oncol* 2008;19:614–22.
7 Lauritzen AD, Rodríguez-Ruiz A, von Euler-Chelpin MC, *et al*. An artificial intelligence-based mammography screening protocol for breast cancer: outcome and radiologist workload. *Radiology* 2022;304:210948:41–9.:.
8 Taylor-Phillips S, Stinton C. Double reading in breast cancer screening: considerations for policy-making. *Br J Radiol* 2020;93:20190610:20190610.:.
9 Byng D, Strauch B, Gnas L, *et al*. AI-based prevention of interval cancers in a national Mammography screening program. *Eur J Radiol* 2022;152:S0720-048X(22)00171-1:110321.:.
10 Lång K, Hofvind S, Rodríguez-Ruiz A, *et al*. Can artificial intelligence reduce the interval cancer rate in mammography screening? *Eur Radiol* 2021;31:5940–7.
11 Freeman K, Geppert J, Stinton C, *et al*. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* 2021;374:n1872:1872.:.
12 Geis JR, Brady AP, Wu CC, *et al*. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Radiology* 2019;293:436–40.
13 Blasimme A, Vayena E. The ethics of AI in biomedical research, patient care, and public health. In: Das S, Pasquale F, Dubber MD, eds. *The Oxford Handbook of Ethics of AI*. New York, NY: Oxford University Press, 2020: 703–18.

14 Rajpurkar P, Chen E, Banerjee O, *et al*. Ai in health and medicine. *Nat Med* 2022;28:31–8.
15 Obermeyer Z, Lee TH. Lost in thought - the limits of the human mind and the future of medicine. *N Engl J Med* 2017;377:1209–11.
16 Huisman M, Ranschaert E, Parker W, *et al*. An international survey on AI in radiology in 1041 radiologists and radiology residents Part 2: expectations, hurdles to implementation, and education. *Eur Radiol* 2021;31:8797–806.
17 Gong B, Nugent JP, Guest W, *et al*. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: anational survey study. *Academic Radiology* 2019;26:566–77.
18 Collado-Mesa F, Alvarez E, Arheart K. The role of artificial intelligence in diagnostic radiology: a survey at a single radiology residency training program. *J Am Coll Radiol* 2018;15:S1546-1440(17)31666-6:1753–7.:.
19 Houssami N, Lee CI, Buist DSM, *et al*. Artificial intelligence for breast cancer screening: opportunity or hype? *The Breast* 2017;36:31–3.
20 Codari M, Melazzini L, Morozov SP, *et al*. Impact of artificial intelligence on radiology: a euroaim survey among members of the European Society of radiology. *Insights Imaging* 2019;10.
21 Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform* 2021;28:e100450.
22 Carter SM, Rogers W, Win KT, *et al*. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast* 2020;49:S0960-9776(19)30564-8:25–32.:.
23 Huisman M, Ranschaert E, Parker W, *et al*. An international survey on AI in radiology in 1,041 radiologists and radiology residents Part 1: fear of replacement, knowledge, and attitude. *Eur Radiol* 2021;31:7058–66.
24 DESI. The Digital economy and society index. European Commission; 2022. Available: https://digital-strategy.ec.europa.eu/en/policies/desi
25 Duffy SW. Mammography screening and research evidence: the Swedish contribution. *Journal of Breast Imaging* 2021;3:637–44.
26 Coppola F, Faggioni L, Regge D, *et al*. Artificial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey. *Radiol Med* 2021;126:63–71.
27 Waymel Q, Badr S, Demondion X, *et al*. Impact of the rise of artificial intelligence in Radiology: What do Radiologists think. *Diagn Interv Imaging* 2019;100:S2211-5684(19)30090-7:327–36.:.
28 Pesapane F, Volonté C, Codari M, *et al*. Artificial intelligence as a medical device in Radiology: Ethical and regulatory issues in Europe and the United States. *Insights Imaging* 2018;9:745–53.
29 Lebovitz S, Lifshitz-Assaf H, Levina N. To engage or not to engage with AI for critical judgments: how professionals deal with opacity when using AI for medical diagnosis. *Organization Science* 2022;33:126–48.
30 Barley SR. Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of Radiology departments. *Adm Sci Q* 1986;31:78–108.

BMJ Health & Care Informatics

Clinical decision support systems to improve drug prescription and therapy optimisation in clinical practice: a scoping review

Lucrezia Greta Armando ,[1] Gianluca Miglio,[1,2] Pierluigi de Cosmo,[3] Clara Cena[1]

To cite: Armando LG, Miglio G, de Cosmo P, *et al*. Clinical decision support systems to improve drug prescription and therapy optimisation in clinical practice: a scoping review. *BMJ Health Care Inform* 2023;**30**:e100683. doi:10.1136/bmjhci-2022-100683

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100683).

Received 22 September 2022
Accepted 12 April 2023

[1]Department of Drug Science and Technology, University of Turin, Torino, Italy
[2]Competence Centre for Scientific Computing, University of Turin, Torino, Italy
[3]Infologic srl, Padova, Italy

Correspondence to
Dr Lucrezia Greta Armando; lucrezia.armando@edu.unito.it

ABSTRACT

Objective Clinical decision support systems (CDSSs) can reduce medical errors increasing drug prescription appropriateness. Deepening knowledge of existing CDSSs could increase their use by healthcare professionals in different settings (ie, hospitals, pharmacies, health research centres) of clinical practice. This review aims to identify the characteristics common to effective studies conducted with CDSSs.

Materials and methods The article sources were Scopus, PubMed, Ovid MEDLINE and Web of Science, queried between January 2017 and January 2022. Inclusion criteria were prospective and retrospective studies that reported original research on CDSSs for clinical practice support; studies should describe a measurable comparison of the intervention or observation conducted with and without the CDSS; article language Italian or English. Reviews and studies with CDSSs used exclusively by patients were excluded. A Microsoft Excel spreadsheet was prepared to extract and summarise data from the included articles.

Results The search resulted in the identification of 2424 articles. After title and abstract screening, 136 studies remained, 42 of which were included for final evaluation. Most of the studies included rule-based CDSSs that are integrated into existing databases with the main purpose of managing disease-related problems. The majority of the selected studies (25 studies; 59.5%) were successful in supporting clinical practice, with most being pre–post intervention studies and involving the presence of a pharmacist.

Discussion and conclusion A number of characteristics have been identified that may help the design of studies feasible to demonstrate the effectiveness of CDSSs. Further studies are needed to encourage CDSS use.

BACKGROUND

Healthcare systems are affected by numerous factors that can reduce quality of care and increase the costs of the services offered. Medication errors are a relevant problem that must be faced with an eye to both patient safety and healthcare-system sustainability. The total costs associated with medication errors in the USA have been estimated at US$42 billion/year and a study has revealed that medication errors during hospital stays may affect up to 6.2% of prescribed medications in the USA and up to 1.5% in the UK.[1 2]

A prescription error may be caused by handwriting problems and poor treatment decisions, potentially leading to the inappropriate use of drugs and harm for patients.[3] Excessive and inappropriate prescriptions can result in severe consequences, such as adverse drug reactions, increased risk of toxicity, prolonged hospital stays, increased antimicrobial resistance, decreased faith in the medical profession and wastage of public funding.[4] This problem is particularly relevant for patients suffering from multiple chronic diseases and requiring the concomitant prescription of different drug classes, a condition that increases the likelihood of medication errors and of potentially inappropriate medications (PIMs) prescription.

Digital technologies[5] including Clinical Decision Support Systems (CDSSs) represent possible strategies for the prevention and reduction of prescription errors. CDSSs consist of digital tools designed to provide interactive computer-based information to assist healthcare professionals in the clinical decision-making process. They were first developed 50 years ago with the aim of promoting optimal problemsolving, decision-making and facilitating the actions of decision-makers as well as making patient data easier to assess. In addition to the support provided to healthcare professionals, CDSSs can produce additional knowledge to guide clinicians by generating new evidence in real time, thus promoting the practice of evidence-based medicine.[6]

Traditional CDSSs consist of a clinical knowledge base, which is the inference engine that combines information from the knowledge base with input data, and of the user interface. In general, it incorporates concepts that

are derived from scientific literature and expert knowledge and should be constantly updated to keep up with new evidence generated in clinical practice.[7] Traditional CDSSs can offer clinicians patient-specific advice based on globally recognised recommendations, as well as increase physician adherence to medical guidelines.

Non-knowledge-based CDSSs are based on artificial intelligence (AI-CDSSs) and have been recently introduced in clinical practice.[8 9] AI-CDSSs still require a data source but leverage AI and machine learning to generate recommendations tailored to patient characteristics. Modern CDSSs are primarily knowledge based since AI-CDSSs require computer-intensive and time-consuming processes and the analysis of a significant amount of data to provide accurate decisions.

The use of these systems has been widely discussed and promoted by healthcare services. They can be used for multiple purposes, including diagnostics, prescription and alarm systems. However, the introduction of CDSSs into all areas of clinical practice still faces several obstacles, including the low ease of system use, negative end-user attitudes towards the system, inaccurate and poor-quality data or documentation, fragmented workflows, financial challenges and an excess of insignificant alerts (alert fatigue).[10 11] New studies should be designed based on the evaluation of previous interventions with CDSSs, regardless of the healthcare setting selected, in order to identify barriers to be overcome for their implementation and key characteristics which proved to generate a positive impact on patient health and on clinicians performances. Although previous studies[12–15] have already estimated the ability of CDSSs to improve healthcare, this kind of evidence has not yet been achieved.

Therefore, the aim of this scoping review is to identify the characteristics of studies in which a CDSS has been effectively implemented in any area of clinical practice producing positive outcomes. Secondary objective is to propose a checklist to be used by healthcare professionals for the implementation of future interventions aimed at demonstrating the effectiveness of CDSSs in improving the quality of care in different settings (ie, hospitals, community pharmacies, general practitioner's (GP) clinics).

MATERIAL AND METHODS

Search strategy

This scoping review was performed according to the guidelines for Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR).[16] The PRISMA-ScR checklist is shown in online supplemental table S1.

Scopus, PubMed, Ovid MEDLINE and Web of Science were searched in order to identify relevant articles. The following filters were applied: article language Italian or English, publication date between January 2017 and January 2022, excluding reviews. Given the recent introduction in clinical practice of CDSSs to improve prescriptive appropriateness, a 5-year period was considered sufficient to identify eligible studies. The final search was conducted on 10 January 2022. The author LGA performed an initial search in Scopus with a combination of the terms "Clinical Decision Support System" and "Inappropriate Prescriptions" to identify relevant keywords. The keywords extracted from the most relevant titles and abstracts were discussed by the authors to select those to be used for the final search. Subsequently, the identified keywords were associated with the Medical Subject Heading terms and approved by all the authors. The final search was conducted by LGA with the query shown in figure 1 and was verified by CC.

(CDSS[MeSH Terms] AND Intervention[MeSH Terms] AND Medication Reconciliation[MeSH Terms])
OR (CDSS[MeSH Terms] AND Intervention[MeSH Terms] AND Appropriate Prescribing[MeSH Terms])
OR (CDSS[MeSH Terms] AND Intervention[MeSH Terms] AND Medication Therapy Management[MeSH Terms])
OR (CDSS[MeSH Terms] AND Intervention[MeSH Terms] AND Medication Errors Screening)
OR (CDSS[MeSH Terms] AND Intervention[MeSH Terms] AND Therapeutic Errors Prevention)
OR (CDSS[MeSH Terms] AND Clinical Study AND Medication Reconciliation[MeSH Terms])
OR (CDSS[MeSH Terms] AND Clinical Study AND Appropriate Prescribing[MeSH Terms])
OR (CDSS[MeSH Terms] AND Clinical Study AND Medication Therapy Management[MeSH Terms])
OR (CDSS[MeSH Terms] AND Clinical Study AND Medication Errors Screening)
OR (CDSS[MeSH Terms] AND Clinical Study AND Therapeutic Errors Prevention)

Figure 1 Final search query.

Eligibility criteria

The question that drove this review was 'Can we learn from previous studies which characteristics and design should have interventions that effectively leverage CDSSs to improve quality of care and prescriptive appropriateness?'. To answer this question, prospective and retrospective studies that reported original research on CDSSs for clinical practice support were identified. Studies including a measurable comparison of the intervention or observation conducted with and without the CDSS were included. Randomised, observational, diagnostic and mixed-method studies were included, while qualitative (survey and semi-structured interview) studies and development reports were excluded. The review does not include studies or documents that describe computerised systems that do not provide decision support, such as electronic health record (EHRs), apps or web-based platforms for therapy self-management. The list of eligibility criteria is given in table 1.

Study selection

The search results were extracted by LGA into a table in Microsoft Excel to remove duplicates. Two authors (LGA and CC) then independently screened study titles and abstracts for inclusion and exclusion criteria. In case of disagreement between LGA and CC, the other authors were asked whether or not to include the study in the next step. Where available, the full texts of potentially relevant articles were screened by LGA and subsequently confirmed by CC. Unavailable full-text articles were defined as publications that could not be accessed either electronically or via a library.

Publications were included if they described a CDSS that was implemented in a real clinical setting and used by healthcare providers to aid decision-making. All systems

Table 1 Inclusion and exclusion criteria for title and abstract screening

	Inclusion	Exclusion
Title and abstract screening	The publication contains research on a CDSS to support clinical decision.	Literature reviews, study protocols, commentaries and editorials were excluded; grey literature was not considered.
	The publication describes the implementation of the CDSS in clinical practice.	The publication has no abstract or full text available.
	The CDSS is used by healthcare professionals to support decision-making.	The publication is written in any language other than English and Italian.
	The publication contains outcomes to measure the effect of the CDSS.	The publication contains a digital tool that does not provide decision support.
	The interventional or observational study analysed includes a comparison between the clinical decision performed with or without the CDSS.	The publication contains a digital tool to be used only by patients and caregivers.
		The publication contains an algorithm or a score not implemented in a computerised system.
		The publication describes telemedicine approaches.

CDSS, Clinical Decision Support Systems.

that analysed patient-specific information to generate case-specific guidance messages through rule-based and algorithm-based software were considered valid, regardless of the targeted assistance (eg, diagnostics tests, chronic disease management, therapy recommendations, drug prescribing, medication reconciliation, medication error detection). Moreover, studies had to report at least one outcome that was capable of measuring the effect of the CDSS on the quality of care provided to patients.

Data extraction

A Microsoft Excel spreadsheet was prepared to extract data from the included articles (online supplemental table S2). All of the authors agreed on what data items to extract to guide the process of result elaboration. Data recorded included: medical area of interest; characteristics of the system and its focus; study setting and design; end user; outcomes measured; study period; sample characteristics; summary of the results.

RESULTS

The search identified 7476 articles eligible for screening. After removing duplicates, 2453 articles were available for title and abstract screening. During title screening, 1975 articles were excluded either because they were of the wrong publication types or lacked a digital tool to support clinical decision; 478 articles were considered to be relevant for abstract screening. This number was further reduced for the reasons given in table 1. After assessing the eligibility of the remaining 136 articles, 42 articles were included in the review. The screening and eligibility-checking process is described in figure 2.

Only seven studies (16.7%) were implemented in more than one setting; 19 (45.2%) were developed in the USA, with the remaining CDSSs being implemented in Canada (4; 9.5%), Australia (3; 7.1%), the Netherlands (3; 7.1%), the United Kingdom (3; 7.1%), China (2; 4.8%), Germany (1; 2.4%), Italy (1; 2.4%), Ireland (1; 2.4%), Norway (1; 2.4%), Austria (1; 2.4%), Switzerland (1; 2.6%), Pakistan (1; 2.4%) and South Korea (1; 2.4%). Table 2 summarises the main characteristics of the studies included in the analysis.

Overview of results

The main setting of the studies analysed was hospital wards, followed by GP clinics and the emergency department. The selected studies focused on the management of various conditions, the most common being the treatment of hospitalised patients and the treatment of children and adolescents. Of the included studies, 40.5% (17 studies) were randomised controlled trials (RCTs), 31.0% (13) before-and-after studies, 23.8% (10) retrospective observational studies and 4.8% (2) non-controlled clinical trials and quasi-experimental design studies.

There were no substantial differences between the number of CDSSs implemented for the management of drug-related problems (22 studies; 52.4%) and that of CDSSs employed to manage problems related to the disease (20; 47.6%). Most of the systems used in the selected studies were knowledge-based CDSSs (35 studies; 83.3%), containing either rules based on globally recognised criteria, such as the Beers criteria and the Screening Tool of Older Persons' Prescriptions (STOPP)[17 18] or rules based on international guidelines. Several platforms for delivering clinical decision support were used, but more than half (22 studies; 52.4%) were CDSSs integrated with existing databases, such as EHRs and/or other hospital electronic devices.

Patient complexity was classified into three levels based on patients' baseline characteristics.

Study description and reported outcomes

The major primary outcome defined by the analysed studies is summarised in table 3. The outcomes are classified according to the level on which they had the greatest

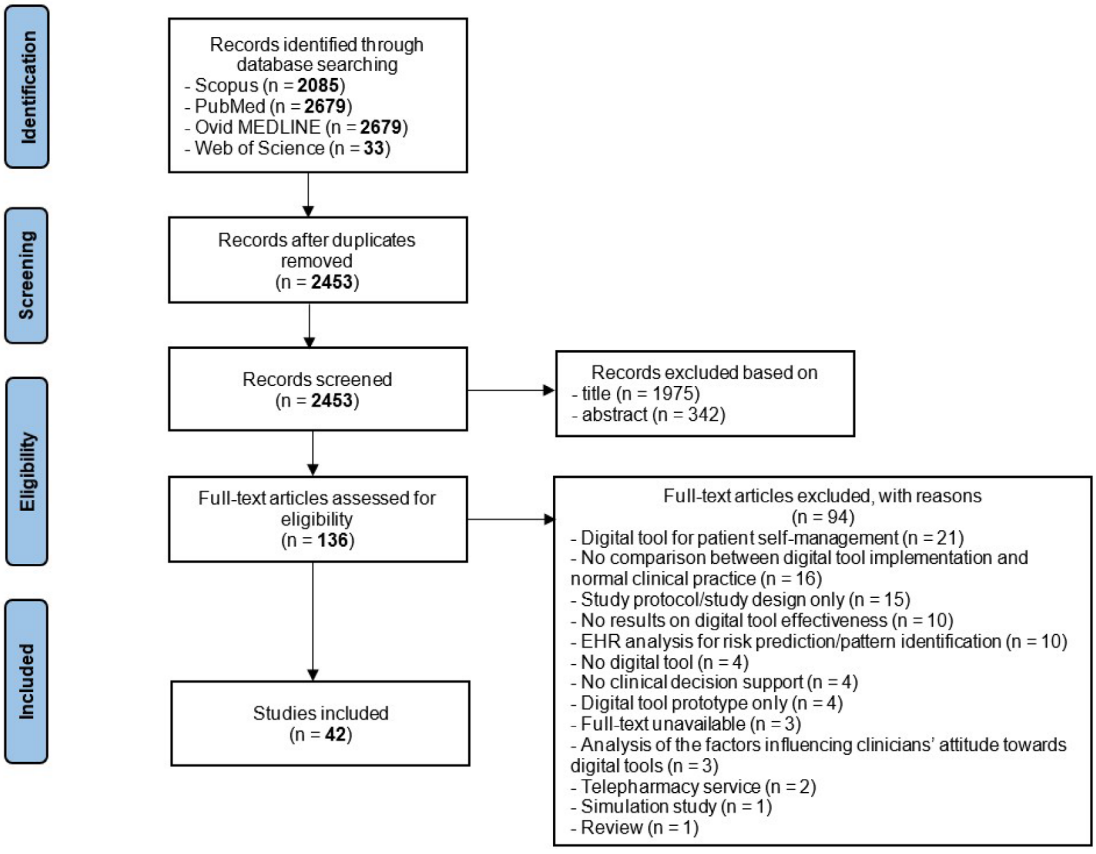

Figure 2 PRISMA flowchart for article selection and review. EHR, electronic health record; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

impact: patient level, clinician level and healthcare setting level.

Twenty-five out of 42 studies achieved their primary outcome with significant differences between the control group and the intervention group, demonstrating the usefulness of CDSSs in improving clinical practice. The online supplemental materials include the characteristics of each study analysed (**S2**) and a figure representing the process for conducting effective studies (**S3**).

Successful studies, that is, those in which the CDSSs were proven to be effective in supporting clinical practice, showed some substantial differences from studies where the CDSSs either failed to support clinical practice (12; 28.6%) or produced uncertain results (5; 11.9%). These latter included studies in which, despite the potential positive effects of CDSSs, the outcomes were not achieved due to study limitations that were highlighted by the authors themselves (ie, short study period, non-homogeneous case–control samples, poorly defined outcomes, non-significant differences between groups).

DISCUSSION

To the best of our knowledge, this is the first scoping review which attempts to identify the characteristics of studies in which different types of CDSSs were used to effectively support clinical decision in different settings. Previous scoping reviews have focused on CDSSs for medication review, rare-disease diagnosis, non-knowledge-based clinical decision support tools and on CDSSs to be used in nursing homes.[6 19–21]

In most of the studies analysed, the implementation of CDSSs in clinical practice improved disease management, increasing the number of PIMs detected, reducing the number of patients who experienced adverse outcomes and enhancing the prescription of appropriate treatments. This aspect is particularly important for certain categories of patients, such as complex patients that suffer from multiple chronic diseases, who often need their (poly)therapy to be reconciled due to the high number of medications that are coprescribed by different specialists. For example, McDonald *et al*[22] have demonstrated that the inclusion of an electronic decision support tool for deprescribing (MedSafer) in primary care increased the proportion of PIMs that were deprescribed at hospital discharge. MedSafer is able to identify inappropriate medications according to the Beers criteria, the STOPP and the Choosing Wisely list[17 18 23] as well as providing tapering instructions for medications such as benzodiazepines. Another study by Fried *et al*[24] has shown that integrating the Tool to Reduce Inappropriate Medication (TRIM) into EHRs was associated with improvements in shared decision-making and reduced medication reconciliation errors. TRIM evaluates prescription appropriateness based on the potential overtreatment of diabetes

Table 2 Characteristics of the studies included in the analysis

Characteristics analysed	Number of studies including the characteristic (%)		
	Total studies (n=42)	Studies with positive results (n=25)	Studies with negative or uncertain results (n=17)
Setting			
Hospital wards	30 (71.4)	19 (76.0)	11 (64.7)
GP clinic	5 (11.9)	1 (4.0)	4 (23.5)
Emergency department	3 (7.1)	1 (4.0)	2 (11.8)
Clinical centre	2 (4.8)	2 (8.0)	0 (0.0)
Community pharmacy	2 (4.8)	2 (8.0)	0 (0.0)
Number of study sites			
Monocentric	26 (61.9)	17 (68.0)	9 (52.9)
Multicentric	14 (33.3)	6 (24.0)	8 (47.1)
NA	2 (4.8)	2 (8.0)	0 (0.0)
Clinical area			
Hospitalised patients	7 (16.7)	4 (16.0)	3 (17.6)
Paediatrics	6 (14.3)	4 (16.0)	2 (11.8)
Infectious diseases	5 (11.9)	4 (16.0)	1 (5.9)
Geriatrics	5 (11.9)	1 (4.0)	4 (23.5)
Chronic non-hospitalised patients	3 (7.1)	3 (12.0)	0 (0.0)
Respiratory diseases	3 (7.1)	2 (8.0)	1 (5.9)
Nephrology	3 (7.1)	2 (8.0)	1 (5.9)
Cardiology	3 (7.1)	1 (4.0)	2 (11.8)
Diabetology	2 (4.8)	1 (4.0)	1 (5.9)
Substance use disorder	2 (4.8)	1 (4.0)	1 (5.9)
Oncology	1 (2.4)	1 (4.0)	0 (0.0)
Haematology disorders	1 (2.4)	1 (4.0)	0 (0.0)
Neurology	1 (2.4)	0 (0.0)	1 (5.9)
Purpose of application			
Disease-related			
Disease treatment and management	16 (38.1)	10 (40.0)	6 (35.3)
Risk assessment of adverse outcomes	3 (7.1)	3 (12.0)	0 (0.0)
Diagnosis	1 (2.4)	1 (4.0)	0 (0.0)
Drug-related			
Medication review	12 (28.6)	6 (24.0)	6 (35.3)
Prescriptive appropriateness	7 (16.7)	4 (16.0)	3 (17.6)
Deprescription	3 (7.1)	1 (4.0)	2 (11.8)
Study design			
RCT	17 (40.5)	5 (20.0)	12 (70.6)
Pre-post intervention study	13 (31.0)	11 (44.0)	2 (11.8)
Retrospective, observational study	10 (23.8)	7 (28.0)	3 (17.6)
Non-controlled intervention study	1 (2.4)	1 (4.0)	0 (0.0)
Quasi experimental design	1 (2.4)	1 (4.0)	0 (0.0)
CDSS characteristics			
Rule-based	22 (52.4)	11 (44.0)	11 (64.7)
Guidelines	13 (31.0)	9 (36.0)	4 (23.5)
AI-based	3 (7.1)	2 (8.0)	1 (5.9)

Continued

Table 2 Continued

Characteristics analysed	Number of studies including the characteristic (%)		
	Total studies (n=42)	Studies with positive results (n=25)	Studies with negative or uncertain results (n=17)
Digital checklist	2 (4.8)	1 (4.0)	1 (5.9)
Predictive models	2 (4.8)	2 (8.0)	0 (0.0)
Platform for CDSS delivery			
Integrated into EHRs	18 (42.9)	11 (44.0)	7 (41.2)
Web-based software	9 (21.4)	6 (24.0)	3 (17.6)
Smartphone-based application	4 (9.5)	2 (8.0)	2 (11.8)
Integrated with CPOE	3 (7.1)	2 (8.0)	1 (5.9)
Integrated into a vital sign monitor	1 (2.4)	0 (0.0)	1 (5.9)
NA	7 (16.7)	4 (16.0)	3 (17.6)
Baseline patient complexity			
High complexity			
Chronic kidney disease	2 (4.8)	2 (8.0)	0 (0.0)
Need for feeding tube	2 (4.8)	2 (8.0)	0 (0.0)
Children	2 (4.8)	2 (8.0)	0 (0.0)
Polymedicated with ≥10 drugs	2 (4.8)	1 (4.0)	1 (5.9)
Need for resuscitation	2 (4.8)	1 (4.0)	1 (5.9)
Therapy with high-risk drugs	1 (2.4)	1 (4.0)	0 (0.0)
Cancer	1 (2.4)	1 (4.0)	0 (0.0)
Medium complexity			
Infectious disease	3 (7.1)	2 (8.0)	1 (5.9)
Opioid use disorder	2 (4.8)	1 (4.0)	1 (5.9)
Need for epidural anaesthesia	1 (2.4)	0 (0.0)	1 (5.9)
Lower complexity			
Unspecified comorbidities	6 (14.3)	2 (8.0)	4 (23.5)
Polymedicated with ≥4 drugs	3 (7.1)	1 (4.0)	2 (11.8)
Asthma	2 (4.8)	2 (8.0)	0 (0.0)
Diabetes	2 (4.8)	1 (4.0)	1 (5.9)
COPD	1 (2.4)	0 (0.0)	1 (5.9)
Adrenal insufficiency	1 (2.4)	0 (0.0)	1 (5.9)
Neuropathy	1 (2.4)	0 (0.0)	1 (5.9)
NA	8 (19.0)	6 (24.0)	2 (11.8)
Duration of the intervention (after CDSS implementation)			
≤6 months	12 (28.6)	9 (36.0)	3 (17.6)
7–12 months	5 (11.9)	5 (11.9)	0 (0.0)
13–18 months	4 (9.5)	1 (4.0)	3 (17.6)
19–24 months	6 (14.3)	3 (12.0)	3 (17.6)
>24 months	7 (16.7)	5 (11.9)	2 (11.8)
NA	8 (19.0)	2 (8.0)	6 (35.3)
CDSS users			
Multidisciplinary team	18 (42.9)	10 (40.0)	8 (47.1)
Clinician	10 (23.8)	4 (16.0)	6 (35.3)
Pharmacist and/or pharmacy technician	7 (16.7)	6 (24.0)	1 (5.9)
GP	3 (7.1)		

Continued

Table 2 Continued

Characteristics analysed	Number of studies including the characteristic (%)		
	Total studies (n=42)	Studies with positive results (n=25)	Studies with negative or uncertain results (n=17)
Researcher	3 (7.1)	3 (12.0)	0 (0.0)
Nurse	1 (2.4)	1 (4.0)	0 (0.0)
Pharmacist participation			
No	24 (57.1)	12 (48.0)	12 (70.6)
Yes	18 (42.9)	13 (52.0)	5 (11.9)

AI, artificial intelligence; CDSS, Clinical Decision Support System; COPD, chronic obstructive pulmonary disease; CPOE, Computerised Physician Order Entry; EHR, electronic health record; GP, general practitioner; NA, not available; RCT, randomised controlled trial.

mellitus and hypertension in the elderly, the Beers and the STOPP criteria, inappropriate renal dosing and patient reports of adverse medication effects.

The main finding of this review is the identification of the characteristics that are most likely associated with positive and negative outcomes, identified by comparing successful and unsuccessful studies. Hospital wards were the most common setting in all studies analysed, although there were substantial differences in the types of patients enrolled: most successful studies first involved the enrolment of hospitalised patients, of children and adolescents and of patients with infectious diseases, while most of the

Table 3 Major primary outcome measures of the analysed studies

Primary outcome measure	Total number of studies including the outcome (%)	Number of studies with positive clinical outcome (%)
Impact on patients		
Number of inappropriate prescriptions	11 (26.2)	5 (45.5)
Resolution rate of medical problems identified	4 (9.5)	3 (75.0)
Risk score assessment	4 (9.5)	2 (50.0)
Number of (re)-hospitalisations	3 (7.1)	2 (66.7)
Asthma control	2 (4.8)	2 (100.0)
Acute kidney injury progression	2 (4.8)	2 (100.0)
Impact on glycaemic control	2 (4.8)	1 (50.0)
Delirium duration and severity	1 (2.4)	0 (0.0)
Feasibility of the intervention and patient satisfaction	1 (2.4)	0 (0.0)
Patient-clinician medication-related communication	1 (2.4)	1 (100.0)
Number of adverse drug events	1 (2.4)	0 (0.0)
Overall studies	**32 (76.2)**	**18 (56.2)**
Impact on clinicians		
Prescription rate of drugs of interest	3 (7.1)	1 (33.3)
Compliance with epidural infusion initiation	1 (2.4)	1 (100.0)
Diagnosis accuracy	1 (2.4)	1 (100.0)
Hypertension recognition	1 (2.4)	1 (100.0)
Time to administration of intravenous antibiotics	1 (2.4)	1 (100.0)
Overall studies	**7 (16.7)**	**5 (71.4)**
Impact on healthcare setting		
Number of appropriate ferritin test orders	1 (2.4)	1 (100.0)
Number of feeding tube-related medication errors	1 (2.4)	1 (100.0)
Percentage of vital signs documented	1 (2.4)	0 (0.0)
Overall studies	**3 (7.1)**	**2 (66.7)**
Total number of studies	**42 (100.0)**	**25 (59.5)**

unsuccessful or inconclusive studies were carried out in geriatric wards. In most successful and unsuccessful studies, CDSSs were intended to be used by multidisciplinary teams operating within a single hospital or clinical centre, underlining the importance of the participation of different healthcare professionals in improving the management of complex patients. The presence of a multidisciplinary team in the clinical decision process facilitated the sharing of information between healthcare professionals; in addition, belonging to a single hospital or clinical centre may have made relationships easier. On the other hand, a large proportion of interventions including multicentre settings proved to be unsuccessful, suggesting that geographical distance may not have favoured multidisciplinary collaboration. Two important differences were found regarding the aim and study design of the studies analysed. First, CDSSs used in successful studies mostly had the aim of managing disease-related problems, whereas the use of CDSSs to support deprescription and/or the appropriate use of drugs was more frequent in unsuccessful and inconclusive studies. Second, most of RCTs produced either unsuccessful or inconclusive studies. This supports the conclusion that case–control studies are likely to fail to demonstrate the efficacy of CDSSs, as it is difficult to enrol comparable samples in terms of patient complexity.

As expected, the use of rule-based CDSSs that were integrated into existing software prevailed with similar proportions in all studies, since these are the simplest and fastest systems to be develop and use.

Baseline patient complexity was a further characteristic that was assessed qualitatively. Patients enrolled in successful studies generally appeared to be more complex at baseline as they had more coprescribed drugs, required enteral nutrition or the prescription of drugs with high risk of interactions or had impaired renal function and infectious diseases. This highlights that the use of CDSSs may especially support the management of complex patients at risk of adverse outcomes. Moreover, optimising the treatment of more complex patients offers greater benefits in terms of both economy and patient well-being, thus improving the quality of care.[25]

The participation of a pharmacist in interventions was also evaluated. Most successful studies included the pharmacist as part of the multidisciplinary team or as the principal investigator, while most of the unsuccessful and uncertain studies did not involve this professional figure; therefore, it is possible to hypothesise that the participation of a pharmacist in interventions could favour more positive outcomes. In support of this hypothesis, numerous studies demonstrated the role of pharmacists in reducing medication errors thanks to their special expertise and in providing education to other healthcare professionals.[26 27]

Finally, education of healthcare professionals and patient engagement were considered. Most successful studies (56.0%) included a preintervention period of education and training for healthcare professionals involved in the use of the CDSS, while only 35.3% of the unsuccessful studies included it; this aspect could, therefore, favour the usability of CDSSs. A general lack of activities to improve patient engagement was observed in all the selected studies: the absence of a summary report for the patient and of follow-up after the intervention in most studies represent a limit that should be overcome in the future by including the level of patient involvement as an outcome.

To evaluate the use of CDSSs at the national level, an assessment of the studies implemented in Italy was made. Despite Italy has a large proportion of elderly suffering from multimorbidity,[28 29] only a few tools have been made available to support clinical decision compared with other countries. Only one Italian study conducted by Moja *et al*[30] proved useful in supporting clinical practice, while three publications were excluded in the last selection phase for the following reasons: in the study conducted by Traina *et al*,[31] the CDSS NavFarma was effectively used to reconcile the therapy of a group of elderly patients without being compared with a control group; in the second excluded study, Cattaneo *et al*[32] used the CDSS INTERcheck to assess the risk of drug–drug interactions and PIMs in patients with COVID-19 at hospital discharge; the last excluded study[33] described the design of a platform (Pneulytics) for the remote monitoring and management of patients with chronic obstructive pulmonary disease.

Based on these findings, the most feasible study design aimed at successfully improving the quality of care with the support of CDSSs gaining significant evidence of outcomes consists in a pre–post intervention study involving hospitalised patients with one or more chronic diseases and a complex situation at baseline, polymedicated and most at risk of adverse outcomes. Considering the length of studies with positive outcomes, at least a 1-year study period including both intervention and preintervention periods should allow differences to be observed in terms of prescriptive appropriateness, frequency and severity of symptoms and, more generally, of disease management. Therefore, enrolled patients should preferably have a life expectancy longer than 1 year to allow for adequate periods of observation before and after CDSS implementation. In order to enable comparison of different studies, authors should identify measurable and quantifiable outcomes at each stage of the study. The ideal CDSS should be easy to use, make information readily available and be integrated into the computerised systems of the healthcare facility where the study is performed, so as to reduce analysis time and the possibility of errors during data transfer. Moreover, studies should include a time for sharing the specific expertise of the different healthcare professionals involved in patient management, including pharmacists, in order to achieve the best possible outcome; active patient engagement in the management of their condition also appears to be associated with better outcomes.

Data on AI-CDSSs are still too limited to make a case for their superiority—or inferiority—over traditional CDSSs.

Strengths and limitations

The main strength of this review is the number of databases queried, along with the inclusion of all types of studies regardless of their focus. This revealed a large number of studies eligible for analysis to identify as many characteristics associated with positive outcome as possible.

The main limitations are the lack of unambiguous taxonomy to describe digital tools that support clinical decision and of recognised recommendations for conducting such studies. For example, some of the studies analysed lacked a description of the data that were entered into the system or did not indicate the end user. The choice to include studies that lacked complete information on the CDSS was made in order to select the largest number of CDSSs that have been used in a real-world healthcare setting.

On one hand, the heterogeneity of the studies has made it difficult for us to compare the different studies and devices (hence, the scoping review), while, on the other, it granted us a global view of the use of CDSSs worldwide.

Another limitation can be found in the absence of a focus on a specific patient category, which made it difficult to assess consistency with previous reviews.

CONCLUSIONS

To sum up, 25 (59.5%) of the selected studies proved effective in supporting clinical practice and improving treatment outcomes in different healthcare scenarios. However, the evidence reported does not allow robust conclusions on the effect of CDSSs in real clinical practice to be drawn, both due to the high variability of the interventions implemented and the limited number of CDSSs found.

From the results of this analysis, an initial version of a checklist was created that could be used to refine the design of studies aimed at evaluating the use of CDSSs:

- Prefer studies with a pre–post intervention scheme.
- Enrol population with complex morbidity and medication regimen at baseline but adequate life expectancy; one hospital setting (one or more wards) should be preferred for subject enrolment.
- Plan interprofessional collaboration and pharmacist involvement.
- Integrate a user-friendly CDSS with the healthcare facility's computerised systems with information sharing capability among healthcare professionals.
- Take into consideration active patient engagement and education of the healthcare professionals involved (contribution still uncertain).

Further research

There is a considerable need for studies that may demonstrate the usefulness of CDSSs in reducing medical errors and improving the quality of care. A possible solution is to promote the use of this checklist to plan studies conducted with CDSSs that may prove effective. Moreover, it would be desirable to validate the checklist and keep it updated according to the latest evidence.

Contributors All authors contributed equally to this work.

Funding This work was supported by funds from the Università degli Studi di Torino, Ricerca Locale Ex 60% 2021 to C. Cena. Grant/award number: not applicable.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iD

Lucrezia Greta Armando http://orcid.org/0000-0001-9395-8047

REFERENCES

1 WHO Global Patient Safety Challenge. Medication without harm. 2017. Available: https://apps.who.int/iris/bitstream/handle/10665/255263/WHO-HIS-SDS-2017.6-eng.id=23E3E58293025B546A38236B7882AFC7?sequence=1
2 Tully MP, Buchan IE. Prescribing errors during Hospital inpatient care: factors influencing identification by pharmacists. *Pharm World Sci* 2009;31:682–8.
3 Rosa MB, Nascimento MMG do, Cirilio PB, *et al*. Electronic prescription: frequency and severity of medication errors. *Rev Assoc Med Bras (1992)* 2019;65:1349–55.
4 Shrestha R, Prajapati S. Assessment of prescription pattern and prescription error in outpatient department at tertiary care district Hospital, central Nepal. *J of Pharm Policy and Pract* 2019;12.
5 Slight SP, Tolley CL, Bates DW, *et al*. Medication errors and adverse drug events in a UK Hospital during the optimisation of electronic prescriptions: a prospective observational study. *Lancet Digit Health* 2019;1:e403–12.
6 Ostropolets A, Zhang L, Hripcsak G. A scoping review of clinical decision support tools that generate new knowledge to support decision making in real time. *J Am Med Inform Assoc* 2020;27:1968–76.
7 Nur Raidah R, Sharifalillah N, Rosma MD. Review on barriers and considerations of clinical decision support system for medication prescribing. 2015 IEEE Student Conference on Research and Development (SCOReD); 2015:489–94
8 Tao L, Zhang C, Zeng L, *et al*. Accuracy and effects of clinical decision support systems integrated with BMJ best practice-aided diagnosis: interrupted time series study. *JMIR Med Inform* 2020;8:e16912.
9 Kessler S, Desai M, McConnell W, *et al*. Economic and utilization outcomes of medication management at a large Medicaid plan with disease management pharmacists using a novel artificial intelligence platform from 2018 to 2019: a retrospective observational study using regression methods. *J Manag Care Spec Pharm* 2021;27:1186–96.
10 Kilsdonk E, Peute LW, Jaspers MWM. Factors influencing implementation success of guideline-based clinical decision support

systems: a systematic review and gaps analysis. *Int J Med Inform* 2017;98:56–64.
11 Sutton RT, Pincock D, Baumgart DC, *et al*. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020;3:17.
12 Kwan JL, Lo L, Ferguson J, *et al*. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. *BMJ* 2020;370:m3216.
13 Taheri Moghadam S, Sadoughi F, Velayati F, *et al*. The effects of clinical decision support system for prescribing medication on patient outcomes and physician practice performance: a systematic review and meta-analysis. *BMC Med Inform Decis Mak* 2021;21:98.
14 El Asmar ML, Dharmayat KI, Vallejo-Vaz AJ, *et al*. Effect of computerised, knowledge-based, clinical decision support systems on patient-reported and clinical outcomes of patients with chronic disease managed in primary care settings: a systematic review. *BMJ Open* 2021;11:e054659.
15 Moja L, Kwag KH, Lytras T, *et al*. Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. *Am J Public Health* 2014;104:e12–22.
16 Tricco AC, Lillie E, Zarin W, *et al*. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018;169:467–73.
17 By the 2019 American Geriatrics Society Beers Criteria® Update Expert Panel. American geriatrics Society 2019 updated AGS beers criteria® for potentially inappropriate medication use in older adults. *J Am Geriatr Soc* 2019;67:674–94.
18 O'Mahony D, O'Sullivan D, Byrne S, *et al*. STOPP/START criteria for potentially inappropriate prescribing in older people: version 2. *Age Ageing* 2015;44:213–8.
19 Damoiseaux-Volman BA, Medlock S, van der Meulen DM, *et al*. Clinical validation of clinical decision support systems for medication review: a scoping review. *Br J Clin Pharmacol* 2022;88:2035–51.
20 Schaaf J, Sedlmayr M, Schaefer J, *et al*. Diagnosis of rare diseases: a scoping review of clinical decision support systems. *Orphanet J Rare Dis* 2020;15:263.
21 Abdellatif A, Bouaud J, Lafuente-Lafuente C, *et al*. Computerized decision support systems for nursing homes: a scoping review. *J Am Med Dir Assoc* 2021;22:984–94.
22 McDonald EG, Wu PE, Rashidi B, *et al*. The medsafer study: a controlled trial of an electronic decision support tool for deprescribing in acute care. *J Am Geriatr Soc* 2019;67:1843–50.
23 Hudzik B, Hudzik M, Polonski L. Choosing wisely: avoiding too much medicine. *Can Fam Physician* 2014;60:873–6,
24 Fried TR, Niehoff KM, Street RL, *et al*. Effect of the tool to reduce inappropriate medications on medication communication and deprescribing. *J Am Geriatr Soc* 2017;65:2265–71.
25 Jaam M, Naseralallah LM, Hussain TA, *et al*. Pharmacist-Led educational interventions provided to healthcare providers to reduce medication errors: a systematic review and meta-analysis. *PLoS One* 2021;16:e0253588.
26 Jokanovic N, Tan EC, Sudhakaran S, *et al*. Pharmacist-Led medication review in community settings: an overview of systematic reviews. *Res Social Adm Pharm* 2017;13:661–85.
27 Brixner D, Biltaji E, Bress A, *et al*. The effect of pharmacogenetic profiling with a clinical decision support tool on healthcare resource utilization and estimated costs in the elderly exposed to polypharmacy. *J Med Econ* 2016;19:213–28.
28 Istat. Rapporto annuale 2018. La situazione del paese. 2018. Available: https://www.istat.it/it/archivio/214230
29 Istat. Le condizioni di salute DELLA popolazione anziana in Italia. 2019. Available: https://www.istat.it/it/archivio/259588
30 Moja L, Polo Friz H, Capobussi M, *et al*. Effectiveness of a hospital-based computerized decision support system on clinician recommendations and patient outcomes: a randomized clinical trial. *JAMA Netw Open* 2019;2:e1917094.
31 Traina S, Armando LG, Diarassouba A, *et al*. Proactive inter-disciplinary CME to improve medication management in the elderly population. *Res Social Adm Pharm* 2021;17:1072–8.
32 Cattaneo D, Pasina L, Maggioni AP, *et al*. Drug-Drug interactions and prescription appropriateness at hospital discharge: experience with COVID-19 patients. *Drugs Aging* 2021;38:341–6.
33 Mongelli M, Orani V, Cambiaso E, *et al*. Challenges and opportunities of iot AD AI in Pneumology. 2020 23rd Euromicro Conference on Digital System Design (DSD); 2020