Accuracy and safety of an autonomous artificial intelligence clinical assistant conducting telemedicine follow-up assessment for cataract surgery

Edward Meinert,[a,b,c,d,*] Madison Milne-Ives,[a,b] Ernest Lim,[e,f,g] Aisling Higham,[e,h] Selina Boege,[a,b] Nick de Pennington,[e] Mamta Bajre,[i] Guy Mole,[e,h,j] Eduardo Normando,[c,**] and Kanmin Xue[h,j,***]

[a]Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK
[b]Centre for Health Technology, School of Nursing and Midwifery, University of Plymouth, Plymouth, UK
[c]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, UK
[d]Faculty of Life Sciences and Medicine, King's College London, London, UK
[e]Ufonia Limited, 104 Gloucester Green, Oxford, UK
[f]Imperial College Healthcare NHS Trust, Western Eye Hospital, London, UK
[g]Department of Computer Science, University of York, York, UK
[h]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[i]Oxford Academic Health Science Network, Oxford Science Park, Robert Robinson Ave, Oxford, UK
[j]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK

eClinicalMedicine 2024;73: 102692
Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102692

Summary

Background Artificial intelligence deployed to triage patients post-cataract surgery could help to identify and prioritise individuals who need clinical input and to expand clinical capacity. This study investigated the accuracy and safety of an autonomous telemedicine call (Dora, version R1) in detecting cataract surgery patients who need further management and compared its performance against ophthalmic specialists.

Methods 225 participants were recruited from two UK public teaching hospitals after routine cataract surgery between 17 September 2021 and 31 January 2022. Eligible patients received a call from Dora R1 to conduct a follow-up assessment approximately 3 weeks post cataract surgery, which was supervised in real-time by an ophthalmologist. The primary analysis compared decisions made independently by Dora R1 and the supervising ophthalmologist about the clinical significance of five symptoms and whether the patient required further review. Secondary analyses used mixed methods to examine Dora R1's usability and acceptability and to assess cost impact compared to standard care. This study is registered with ClinicalTrials.gov (NCT05213390) and ISRCTN (16038063).

Findings 202 patients were included in the analysis, with data collection completed on 23 March 2022. Dora R1 demonstrated an overall outcome sensitivity of 94% and specificity of 86% and showed moderate to strong agreement (kappa: 0.758–0.970) with clinicians in all parameters. Safety was validated by assessing subsequent outcomes: 11 of the 117 patients (9%) recommended for discharge by Dora R1 had unexpected management changes, but all were also recommended for discharge by the supervising clinician. Four patients were recommended for discharge by Dora R1 but not the clinician; none required further review on callback. Acceptability, from interviews with 20 participants, was generally good in routine circumstances but patients were concerned about the lack of a 'human element' in cases with complications. Feasibility was demonstrated by the high proportion of calls completed autonomously (195/202, 96.5%). Staff cost benefits for Dora R1 compared to standard care were £35.18 per patient.

Interpretation The composite of mixed methods analysis provides preliminary evidence for the safety, acceptability, feasibility, and cost benefits for clinical adoption of an artificial intelligence conversational agent, Dora R1, to conduct follow-up assessment post-cataract surgery. Further evaluation in real-world implementation should be conducted to provide additional evidence around safety and effectiveness in a larger sample from a more diverse set of Trusts.

*Corresponding author. DEPTH AI Lab, Translational and Clinical Research Institute, Newcastle University, Campus of Ageing and Vitality, Westgate Road, Newcastle upon Tyne, NE4 5PL, UK.
**Corresponding author. Imperial College Opthalmology Research Group, Imperial College London, Western Eye Hospital, London, NW1 5QH, UK.
***Corresponding author. Nuffield Department of Clinical Neurosciences, University of Oxford, Level 6 West Wing, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK.
E-mail addresses: edward.meinert@newcastle.ac.uk (E. Meinert), e.normando@imperial.ac.uk (E. Normando), kanmin.xue@eye.ox.ac.uk (K. Xue).

Funding This manuscript is independent research funded by the National Institute for Health Research and NHSX (Artificial Intelligence in Health and Care Award, AI_AWARD01852).

Keywords: Artificial intelligence; Natural language processing; Cataract; Telemedicine; Digital health; Clinical study

Research in context

Evidence before this study

Prior to undertaking this study, the authors conducted a systematic review of the effectiveness of artificial intelligence conversational agents in healthcare. We searched PubMed, Medline (Ovid), EMBASE, CINAHL, Web of Science, and the ACM Digital Library on November 29, 2019 using a combination of several Medical Subject Heading (MeSH) terms and keywords relating to conversational agents, health application, and outcome assessment. We included 31 studies (published in English since 2008) with a variety of study designs that evaluated at least one unconstrained natural language processing conversational agent developed for use in healthcare. A variety of types of conversational agents were identified and had overall mixed to positive evidence of effectiveness, usability, and satisfaction, but the study quality was poor to moderate. Key barriers reported by users included difficulties being understood by the conversational agent and forming personal connections with it and repetition and lack of interactivity. There were a variety of definitions of effectiveness in the studies that hindered comparability. A more recent systematic review of conversational agents in healthcare conducted by Li and colleagues in 2023 found similar results: evidence for feasibility and acceptability, with varying levels of effectiveness on specific outcomes; however, a third of the studies had a high risk of bias.

Added value of this study

This study found that an artificial intelligence conversational agent could conduct a follow-up assessment post-cataract surgery and make symptom and care management decisions comparable to a human ophthalmologist. It demonstrated that such a tool was generally usable and acceptable to patients but highlighted that surgical outcome was a key factor influencing patients' attitudes towards an automated follow-up assessment.

Implications of all the available evidence

Our findings support the potential for artificial intelligence conversational agents to automate routine, low-skill healthcare tasks in follow-up assessment post-cataract surgery. They also emphasise the importance of implementing such automation to ensure that clinical resources can be provided in a timely manner to patients with the greatest need, rather than replacing patient–clinician interaction points.

Introduction

Demands on the United Kingdom's (UK) National Health Service (NHS) continue to rise while resources remain limited.[1] Digital technologies have the potential to help address workforce constraints[2] by automating repetitive, lower-skill tasks. Routine 'high volume, low complexity' clinical pathways account for many waiting list backlogs[3] and offer an excellent opportunity for improved efficiency. Cataract surgery is the most common surgery in the UK[4–8] and has low complication rates (1–2%[9–11]). This makes its follow-up assessment an ideal target for automation, which could provide a safety net for patients and free up clinicians' time for skill-demanding tasks.[12] In the UK, the most common follow-up practice is face-to-face (F2F) review, usually at 4 weeks; however, some NHS Trusts have adopted telephone follow-up and found it to be a viable option (although the rates of telephone follow-up are poorly reported) and a number of Trusts only review if symptoms dictate (e.g. patient-initiated follow up).[13] Artificial intelligence conversational agents have the potential to go a step further in freeing clinician time and enabling the prioritisation of patients with complications.

Ufonia Limited (Oxford, UK) developed a voice-based telephone conversational assistant ("Dora R1") to deliver post-cataract surgery follow-up. Dora R1 uses speech transcription, natural language understanding, a machine learning conversation model, and speech generation to ask patients symptom-based questions (Fig. 1).[14] This study aimed to build on preliminary evidence of acceptability[15,16] and evaluate the accuracy, safety, and potential for adoption into routine clinical care of the first version (R1) of Dora to gain CE mark approval. The primary objective was to examine Dora R1's ability to accurately detect whether a patient requires further assessment by a human clinician. Other factors relating to Dora R1's adoption, including patients' perceptions and implementation costs, were also assessed to determine the strength of evidence for broader implementation.

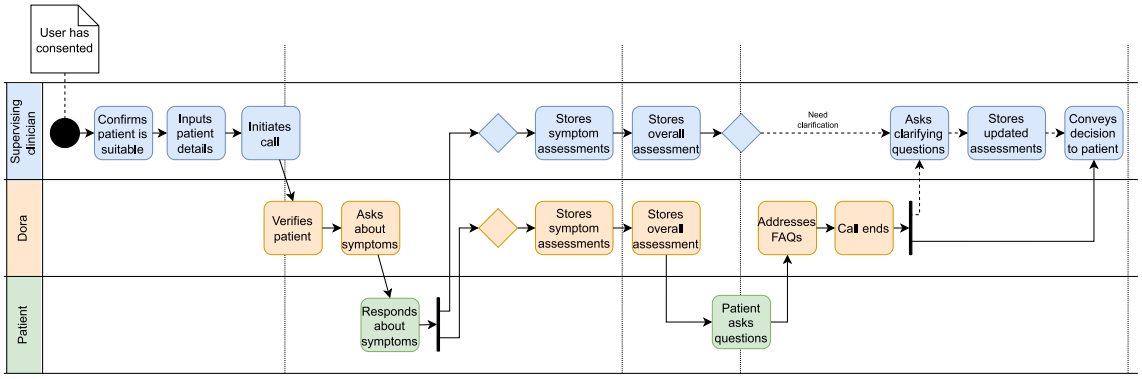

Fig. 1: Flow diagram of Dora R1's call with patients undergoing routine cataract surgery.

Methods

Study design

A mixed-methods implementation science approach[17] was used to evaluate Dora R1's potential impact as a postoperative assessment of cataract surgery patients (Fig. 2).

Ethics statement

Research ethics approval was obtained from the Health Research Authority (21/PR/0767) and the University of Plymouth's Faculty Research Ethics and Integrity Committee (2863). The trial was registered on ClinicalTrials.gov (NCT05213390) and ISRCTN (16038063). The Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI)[18] and Standards for Reporting Qualitative Research (SRQR)[19] checklists were used to ensure comprehensive reporting (Supplementary Tables S1 and S2).

Research participants, recruitment, and procedure

Patients undergoing cataract surgery were recruited over five months (09.2021–01.2022) from two clinical sites: the Imperial College Healthcare NHS Trust (London)

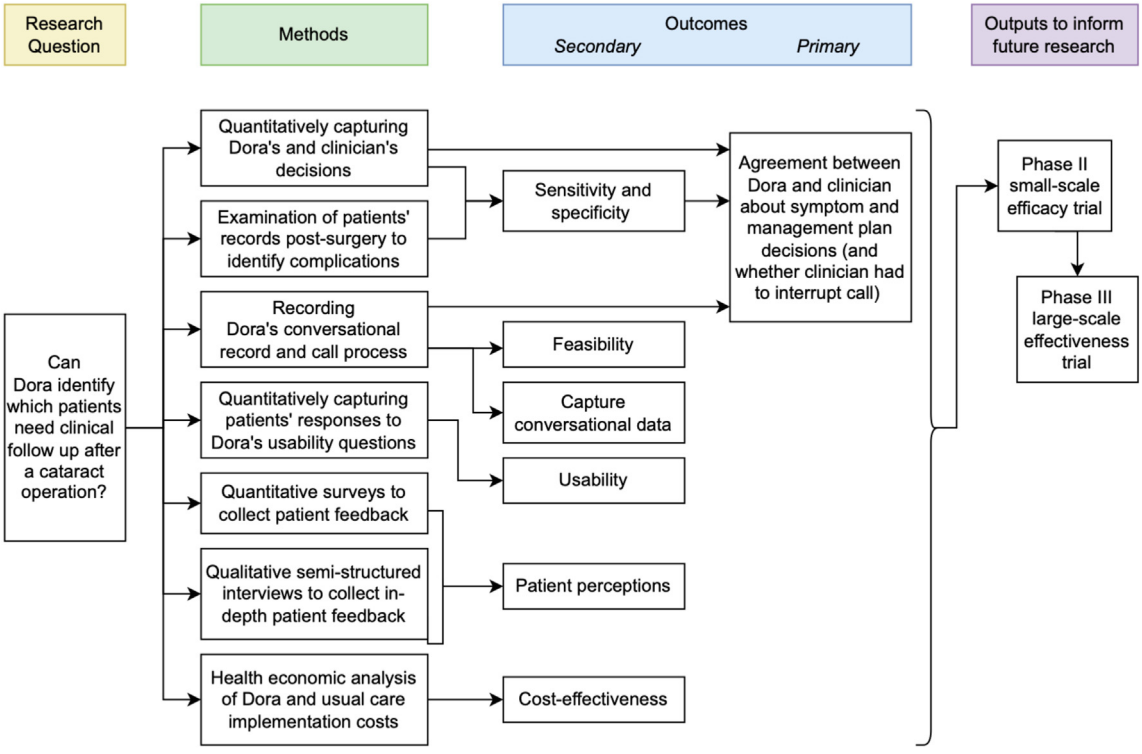

Fig. 2: Logic diagram of the clinical study.

and Oxford University Hospitals NHS Foundation Trust (Oxford). These two sites represent different populations–ethnically diverse and urban in London and primarily white British and more rural in Oxford. All patients who met the eligibility criteria were invited to participate by telephone by research nurses or in pre-assessment sites and post-operative discharge lounges. There were two stages of screening during enrolment: the first stage screened clinical notes to determine potential eligibility and the second involved contacting potentially eligible patients by phone to confirm eligibility and consent them for the study. Informed consent was required to participate; it was collected (either written in person or verbally over the phone) by Research Nurses at the two sites. Consent forms were reviewed and signed by the Principal Investigator at each site.

Participants were eligible for inclusion if they were a) aged 18 years or older, b) willing and able to provide consent, c) on the waiting list for routine cataract surgery, and d) had no history or presence of significant ocular comorbidities that would be expected to alter the risks of cataract surgery or normal post-operative follow-up schedule. Note that significant ocular comorbidities do not include stable, chronic, or inactive ocular conditions such as amblyopia, drop-controlled stable glaucoma or ocular hypertension, previous squint surgery, inactive macular pathology, previous refractive surgery, or previous vitreoretinal surgery with stable retina. Individuals were excluded if they a) had any condition that could preclude the ability to comply with the study or follow-up procedures, b) were having cataract surgery as part of a combined procedure with other ocular surgery, c) had ocular or systemic uncontrolled disease or a history of current or severe, unstable or uncontrolled systemic disease (unless deemed not clinically significant by the Investigator and Sponsor), d) were involved in current research or have been involved in any research within 2 months from enrolment, or e) had cognitive difficulties, hearing impairment or did not speak fluent English.

All eligible patients received a postoperative call from Dora R1 approximately 3–4 weeks post-surgery in addition to their default care pathway. The standard care pathway at Oxford is that patients without significant ocular comorbidities and have routine surgery are discharged on the day of surgery and advised to attend their community optometrist after 6 weeks (Fig. 3). At Imperial, patients typically received face-to-face follow up at 4 weeks after surgery. The 3–4 week time point for the Dora R1 call was chosen as this is prior to the standard 4-week follow-up at Imperial.

Dora R1 calls were supervised in real-time by one of two ophthalmologists who could take over if needed. Dora R1 and the supervising ophthalmologists independently made decisions about the clinical significance of five symptoms and the overall call outcome (either 'recommend discharge' or 'recommend further review'; Supplementary Tables S3 and S4, Supplementary Figure S1). 'Recommend discharge' refers to discharge from hospital-based ophthalmology services; typically, all patients are advised to see their community optometrist for a sight test at 6 weeks. All patients recommended for review by Dora R1 received an immediate call-back by a clinician to determine next steps, which mirrors the real-world operational model.

All patients were subsequently contacted to complete a usability survey and a subset were invited to qualitative semi-structured interviews using stratified random sampling (02–03.2022). Primary and matched backup lists were created by randomly selecting a white male and female participant from each income bracket (n = 12) and 8 patients of non-white ethnicity (there were insufficient participants to select from each income bracket).

Two ophthalmologists conducted all of the call supervision to make independent decisions about symptoms and outcome and to identify errors; one for Oxford patients (AH) and one for Imperial patients (EL). AH is a Fellow of the Royal College of Ophthalmologists with over five years ophthalmology subspecialty training experience. EL is an ophthalmologist with three years of ophthalmology subspecialty training experience. The supervising ophthalmologists were blinded to all of Dora R1's decisions. The supervising ophthalmologist's decisions were the reference standard that Dora was compared against, as this is how standard care would be delivered, and were the decisions that determined patients' subsequent care.

Data collection

The software automatically recorded Dora R1's and the supervising ophthalmologist's decisions (primary outcome was agreement). Clinical data was collected from patients' electronic health record (EHR) up to 3 months postoperatively to capture unplanned attendances or unexpected management changes (UMC). UMC were defined as 1) a deviation from the eye drop taper plan prescribed on the day of surgery for the antibiotic, steroid, or NSAID drop, 2) addition of an eye drop excluding artificial tears, 3) performance of a procedure excluding suture removal, or 4) additional clinic review required. For purposes of analysis we divided events into before and after 2 weeks to capture key acute presentations (e.g. endophthalmitis) related to the assessment as well as delayed concerns (e.g. cystoid macular oedema, rebound uveitis). Demographic and usability data were collected via telephone or online Qualtrics surveys[20] and through patients' electronic health records (EHR) (10.2021–03.2022). Supplementary Text 1 outlines the semi-structured interview topic guide, which was developed based on the Theoretical Framework of Acceptability (TFA).[21] The participant flow is shown in Supplementary Figure S2 and data collected are detailed in

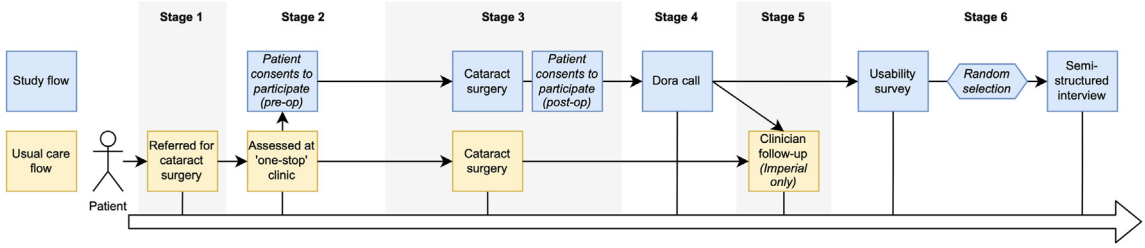

Fig. 3: Participant flow diagram.

Supplementary Table S5. Full details of outcomes are included in the published protocol.[12]

Data analysis

Quantitative data was analysed using Stata (Release 17).[22] The primary analysis calculated inter-rater reliability (Cohen's kappa) between Dora R1 and the clinician, accuracy, sensitivity, specificity, and area under the curve for the overall outcome and each symptom decision. Both Cohen's kappa and sensitivity/specificity metrics were included to provide a comprehensive evaluation of the performance of the conversational agent compared to the supervising ophthalmologist. While sensitivity and specificity measure the accuracy of correct positive and negative classifications, respectively, the kappa statistic accounts for the agreement between observed and expected classifications, correcting for chance agreement. By incorporating both types of metrics, the analysis aimed to assess not only the model's overall accuracy but also its agreement beyond what could be expected by random chance, providing a more nuanced and robust evaluation of its performance. Secondary analyses included calculating the same metrics comparing Dora R1's overall decision with unplanned hospital review and unexpected management changes, using descriptive statistics to examine feasibility and usability (System Usability Scale (SUS total score),[23] Telehealth Usability Questionnaire (TUQ average score),[24] and Net Promoter Score (NPS value)[25]), and conducting exploratory analyses to identify any associations between demographic characteristics and usability.

Missing data was excluded from statistical analyses. A sensitivity analysis was conducted to assess the reliability of the findings. A Little's Missing Completely at Random (MCAR) test was conducted to determine whether data was missing completely at random (missingness independent of observed and unobserved data) and a Mann–Whitney U test was used to compare the differences between two groups for missing value patterns to assess whether data was missing at random (MAR) or missing not at random (MNAR). A Mann–Whitney U test is appropriate for comparing the differences between two groups for missing value patterns when sample sizes are small. Fully Conditional Specification (FCS) imputation was conducted to mitigate potential bias for missing data not at random. This method was selected because it operates within a multiple imputation framework and has a flexible modelling approach that enables individual models to be built based on observed data and accommodate complex missing data patterns to develop accurate imputations.

A codebook approach[26,27] was used to conduct a theoretical thematic analysis of the interviews in QSR NVivo 12.[28] The codebook was based on the TFA constructs: affective attitude, burden, perceived effectiveness, ethicality, intervention coherence, opportunity costs, and self-efficacy.[21] Within this structure, codes were inductively generated from the transcripts by one author.[29] A second author independently coded the transcripts using this codebook, adding additional codes where needed. Both authors independently developed thematic maps,[30] which were compared, discussed, and consolidated.

A cost analysis compared the direct costs of face-to-face (F2F) follow-up at Imperial with Dora R1 (in Oxford, patients do not have routine postoperative follow-up). Assumptions included annual costs for various healthcare professionals[31] and the duration of F2F follow-up appointments (estimated at 30 min).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Overview

Between 17 September 2021 and 31 January 2022, 767 patients were identified as eligible and 225 (30%) provided informed consent. 202 were included in the analysis; 180 completed surveys and 20 participated in semi-structured interviews (Fig. 4, Supplementary Table S6). Reasons for exclusion in the first stage of screening (notes review) were not reported across sites due to varying cadence and methods of screening related to EHR and trial operational processes; the majority were excluded because they would have already had a clinic appointment before the Dora call or because they had significant ocular comorbidity.

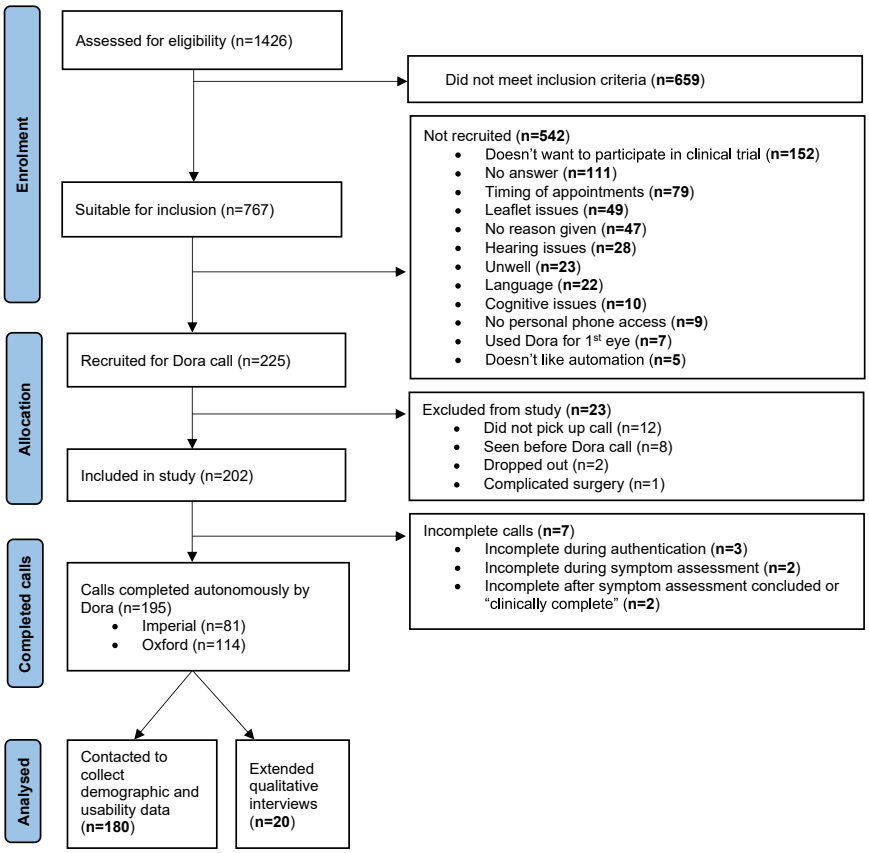

Fig. 4: Flowchart for study recruitment.

Reasons for exclusion at the final screening stage and after recruitment are included in the flow diagram. Across the two sites, there was a mean of 29 days between surgery and Dora calls (SD = 8.13).

Participant characteristics

Demographic characteristics are summarised in Table 1.

Agreement

The primary outcome was the agreement between Dora R1 and the supervising clinician on five key symptoms[14,32] and overall care management decisions. The Kappa statistics indicated a moderate-to-strong level of agreement (Table 2).[33] Dora R1's ability to differentiate between patients experiencing a symptom or not was high (accuracy: 97–99%)[34] and slightly lower for overall outcome decision (accuracy: 89%). Sensitivity and specificity were also high, with the exception of the sensitivity of the decision around vision issues (67%; explored further in the discussion). Out of 948 symptom assessments, Dora R1 had 12 (1.3%) errors and 7 (0.7%) false-negatives (Supplementary Table S7).

Missing data was omitted from the statistical analyses; for Dora and supervisor individual symptom and outcome decisions, missing data ranged from 0 to 12 observations out of 199 potential data points (Table 3). Missing data occurred when Dora or the supervising clinician could not hear or understand the patient's response to a question. A Little's MCAR test was significant (χ^2 (139) = 426.89, $p < 0.001$), indicating that the missingness in the data was not completely random. The subsequent Mann–Whitney U test found a significant difference in the proportions of missing data between Dora R1 and the supervising clinician ($z = -2.741$, $p = 0.004$), with Dora R1 having a higher probability of more missing data (mean rank = 9.3, compared to 3.7 for clinician; Supplementary Figure S3), indicating that data was not missing at random. Five imputations were created using the FCS imputation specification imputation method; the scale variables were modelled using the Predictive Mean Matching technique, with five closest predictions. The performance of the imputed and original analysis data was similar, although the imputed data performed slightly worse for vision issues and new floaters (Supplementary Table S8).

Demographics	Oxford	Imperial	Overall
Number of study participants	119	83	202
Age[a]			
Age–median (range)	76 (42–94)	73 (55–92)	74 (42–94)
Age–IQR	69.5–82	66.5–78.5	68–80
Age–mean (SD)	75 (9.51)	72 (7.96)	74 (8.96)
First or second cataract surgery[a]			
First eye (%)	65 (55)	48 (58)	113 (56)
Second eye (%)	54 (45)	35 (42)	89 (44)
Gender[a]			
Male (%)	52 (44)	34 (41)	86 (43)
Female (%)	67 (56)	49 (59)	116 (57)
Ethnicity[b]			
Total number identified	119	81	200
White (%)	107 (90)	53 (65)	160 (80)
Asian (%)	3 (3)	9 (11)	12 (6)
Black (%)	0 (0)	11 (14)	11 (6)
Other (%)	9 (8)	8 (10)	17 (9)
Income[c]			
Total number identified	88	59	147
≤£19,999/yr (%)	16 (18)	12 (20)	28 (19)
£20,000–29,999/yr (%)	13 (15)	11 (19)	24 (16)
£30,000–39,999/yr (%)	3 (3)	2 (3)	5 (3)
£40,000–49,999/yr (%)	3 (3)	2 (3)	5 (3)
£50,000–69,999/yr (%)	6 (7)	3 (5)	9 (6)
>£70,000/yr (%)	3 (3)	3 (5)	6 (4)
Undefined (%)	44 (50)	26 (44)	73 (50)
Education level[c]			
Total number identified	81	50	131
Lower than bachelor's degree (%)	52 (64)	36 (72)	88 (67)
Bachelor's degree or higher (%)	29 (36)	14 (28)	43 (33)

[a]Data retrieved from patients' electronic health record (EHR). [b]Data retrieved from patients' electronic health record (EHR) and supplemented with self-report questionnaire data if it provided more detailed information. [c]Data collected from self-report questionnaire.

Table 1: **Participant demographic information.**

Dora R1 recommended discharge for 4 patients whom the clinician recommended review (2%, Table 4), none of whom required clinical review on clinician call-back (Supplementary Table S9). Details about the concordance between Dora R1 and the supervisor for each symptom are included in Supplementary Table S10. Dora R1 recommended review for 18 patients whom the clinician discharged (9%, Supplementary Table S11). Disagreements were primarily due to misunderstanding responses or transcription errors, particularly when patients responded with a combination of positives and negatives (e.g. "no, yes").

Safety

We present outcomes from patients recommended for discharge by Dora R1, as these are the patients who would be discharged without further clinician assessment in a real-world setting. Of these 117 patients, 71 (65%) were seen for planned review (48 within two weeks post-surgery; Table 5), but of the patients discharged by Dora R1, only 9% had any symptoms that required an unexpected management change. Five patients had an unplanned attendance: 2 were asymptomatic but had missed their planned review, 2 had new symptoms more than a month post call, and 1 was referred back from community optometry but had no acute findings in hospital. There was only 1 unplanned attendance within 2 weeks of the Dora R1 call (one of the asymptomatic patients). Approximately 10% (11/113) of patients recommended for discharge by Dora R1 and the supervising clinician experienced unexpected management changes, only 1 of which was an unplanned emergency review (a patient 44 days after the Dora R1 call with a diagnosis of rebound anterior uveitis). Supplementary Table S12 includes further clinical details.

Feasibility, acceptability and usability

Efficiency

The supervisor recommended review for a third of patients (64/195). Of these, 36% (23/64) were

Decision (n)	Accuracy (%) [95% CI]	Sensitivity (%) [95% CI]	Specificity (%) [95% CI]	Kappa [95% CI]	p-value
Redness (191)	99.48 [97.12–99.99]	100 [15.81–100.00][a]	99.47 [97.09–99.99]	0.798 [0.410–1.000]	<0.0001
Pain (189)	98.41 [95.43–99.67]	100 [59.04–100.00][a]	98.35 [95.26–99.66]	0.816 [0.612–1.000]	<0.0001
Vision issue (187)	97.86 [94.61–99.41]	66.67 [34.89–90.08]	100 [97.91–100.00][a]	0.789 [0.589–0.989]	<0.0001
New floaters (188)	98.40 [95.41–99.67]	86.96 [66.41–97.22]	100 [97.79–100.00][a]	0.921 [0.833–1.000]	<0.0001
Flashing lights (193)	99.48 [97.15–99.99]	100 [81.47–100.00][a]	99.43 [96.86–99.99]	0.970 [0.912–1.000]	<0.0001
Outcome (195)	88.72 [83.42–92.79]	93.75 [84.76–98.27]	86.26 [79.16–91.56]	0.758 [0.664–0.852]	<0.0001

Note: Accuracy = (TP + TN)/(TP + TN + FP + FN), Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP), p-value was calculated based on the Kappa statistic. (n) refers to the number of decisions made by both Dora and the supervising clinician for each symptom and the overall outcome. [a]One-side, 97.5% confidence interval.

Table 2: **Agreement between Dora R1 and supervising clinician in identifying 5 key symptoms and overall clinical outcome.**

recommended for a F2F assessment on callback and 33% (21/64) required a management change when seen. Overall, only 12% of the cohort (23/197) required F2F review by a clinician.

Timing

Dora R1 completed 96.5% of calls autonomously (195/202), with a mean call length of 7 m 25s (n = 186, SD = 2 m 07s). Length of time per symptom is reported in Supplementary Table S13.

Acceptability and patient experiences

Thematic analysis generated three main themes around acceptability (Fig. 5). The first theme found that the intervention was perceived as an acceptable tool with potential benefits for patients who had not experienced any complications with their surgery and who did not have any concerns. Before the call, a couple participants worried it might be frustrating and a few were interested, but most had no expectations. After the call, most interviewees' attitudes were neutral to positive; some found it "*fine,*" others felt comfortable, appreciated the straight-forward approach, and "*enjoyed talking with Dora.*" Many were confident in Dora R1's ability, with one caveat: around half the interviewees had concerns about using Dora R1 for patients with complications.

Despite these concerns (discussed in Themes 2 and 3), many participants felt that Dora R1 could benefit patients and clinicians by increasing convenience, saving time and costs, and providing reassurance by ensuring that all patients receive follow-up. A few participants even highlighted areas where Dora R1 may be superior to a human: it "*gave [them] time to take a breath*" and think without being "*conscious that [the doctor is] busy and you don't want to look stupid*" and it wouldn't have "*off-days.*" This aligns with quantitative data collected during the call about how likely patients would be to recommend Dora R1 (Net Promoter Score, NPS[25]). The mean individual response was 8.59/10 (*SD* = 2.05) and the overall NPS score, which can range from −100 to 100,[25] was 51.06. Scores over 50 are generally considered to be above average.[35]

Interviewees expressed approximately equal preferences for F2F or Dora R1 consultations, with slightly more favouring F2F. Some indicated that it would depend on the situation or that they would prefer a F2F consultation but "*had no objection to [Dora].*" There was almost universal willingness to use Dora R1 again; only one participant was steadfastly against it, although a few others expressed some reluctance: "*well, if I have to use it, I have to use it.*"

The second main theme revolved around concerns about Dora R1's ability to manage complicated or emotional situations and the importance of the human element in care. Many patients found Dora R1 impersonal, with some saying that they felt "*a bit remote and lonely talking to a machine*" and others disliking the "*mechanical voice*". Some participants felt limited by Dora R1's questions and the lack of a "*two-way exchange*" through which patients could ask questions or expand on their answers. One of the biggest concerns for Dora R1 compared to a clinician was the potential to miss nuances and non-verbal information: "*closed questions… will only get a certain answer. That may not be… the whole picture, but a computer won't pick up on that.*" Many participants were "*happy to talk to AI, if there's nothing wrong,*" but "*would not feel reassured*" if they had concerns. Several stressed the importance of "*a two-step*

Decision	Decider	Missing data points (n)	Available data points (n)	Total possible data points (n)	Missing (%)
Redness	Supervisor	1	198	199	0.50%
	Dora R1	8	191	199	4.02%
Pain	Supervisor	0	199	199	0.00%
	Dora R1	10	189	199	5.03%
Vision issue	Supervisor	4	195	199	2.01%
	Dora R1	12	187	199	6.03%
New floaters	Supervisor	3	196	199	1.51%
	Dora R1	11	188	199	5.53%
Flashing lights	Supervisor	2	197	199	1.01%
	Dora R1	6	193	199	3.02%
Outcome	Supervisor	4	195	199	2.01%
	Dora R1	4	195	199	2.01%

Table 3: **Missing data.**

	Supervisor decision		
Dora R1 decision	Discharge (%)	Review (%)	Total (%)
Discharge (%)	113 (58)	4 (2)	117 (60)
Review (%)	18 (9)	60 (31)	78 (40)
Total (%)	131 (67)	64 (33)	195 (100)

Table 4: **Concordance between Dora R1 and supervisor outcomes.**

process" whereby patients would be called by Dora R1 but *"if [they] have any concerns at all, [be put] through to somebody [they] can speak to."*

The third theme focused on usability and accessibility. Almost all interviewees (mean age = 73.75 years, SD = 8.42, range: 56–90 years) found Dora R1 easy to use and that *"[they] could understand everything it was saying"*. This was in line with the quantitative assessment, which found a mean SUS score of 77.76/100 (*SD* = 17.55) and TUQ score of 3.79/5 (*SD* = 0.89). An SUS score of 77.76 equates to a 'B' on the Sauro-Lewis curved grading scale and is considered relatively good.[23] Usability depended to some degree on users' efforts to provide simple answers to avoid confusing Dora R1 and potential accessibility barriers such as hearing, understanding Dora R1's accent, and *"older people… [finding] it quite difficult to communicate with a computer"* were raised. The intervention was also potentially unsuitable for people with autism; one participant with autism *"was extremely distressed by the whole experience"* because they *"find it… difficult to process information [without] the opportunity to ask relevant questions"* and were not confident in Dora R1's *"ability to identify… a problem that needs further investigation."* Another participant felt their brother, who was *"dyslexic [and] mildly autistic"* would struggle with Dora R1. To improve usability and accessibility, participants suggested further training so Dora R1 can better understand more detailed responses, clarifying the objective of the call, providing questions in advance or an alternative paper version, and facilitating scheduling and reminders.

Exploratory analysis of associations between usability and demographics

As a Shapiro–Wilk W test for normality found that none of the major variables included in these analyses were normally distributed (SUS: W = 0.96, p = 0.00051; TUQ: W = 0.96, p = 0.0018; NPS: W = 0.82, p < 0.0001), non-parametric alternatives to t-tests and Pearson's correlation analyses were conducted. The only statistically significant demographic differences in usability and acceptability scores were in the Wilcoxon rank-sum test comparisons of mean TUQ score and NPS score by education level (TUQ: z (106) = 3.040, p = 0.0024, TUQ: z (121) = 2.621, p = 0.0088)–participants with a Bachelor's degree or higher had significantly lower TUQ scores (TUQ: rank sum = 1,370, NPS: rank sum = 1941.5) than participants those without (TUQ: rank sum = 4,301, NPS: rank sum = 5439.5)–and in the Kruskal–Wallis Equality-of-population test comparison of NPS score by household income (χ^2 (6) = 19.636, p = 0.0032). The difference in NPS score by income was mainly driven

	Oxford	Imperial	All sites
Participants 'recommended discharge' by Dora R1 (n)	80	37	117
Numbers seen for any reviews in 3 months (%)	39 (49%)	37 (100%)	76 (65%)
Planned review			
Number of planned reviews in <2 weeks (% of all reviews)	14 (36%)	34 (92%)	48 (63%)
Of planned reviews, unexpected management changes (% of all recommended for discharge)[a]	3 (4%)	7 (19%)	10 (9%)
Unplanned reviews near Dora R1 calls (Within 2 weeks)			
Unplanned reviews <2 weeks after Dora R1 calla	0	1	1
Of these, had subsequent unexpected management changes	0	0	0
Unplanned reviews (After 2 weeks)			
All unplanned review up to 3 months after Dora R1 call (% of all calls)a	2 (3%)	3 (8%)	5 (4%)
Of all unplanned reviews, had subsequent management change (% of all recommended discharge)	0	1	1

[a]Case-by-case breakdown provided in Supplementary Table S12.

Table 5: **Numbers of participants 'recommended discharge' by Dora R1 with subsequent planned or unplanned reviews or unexpected management change.**

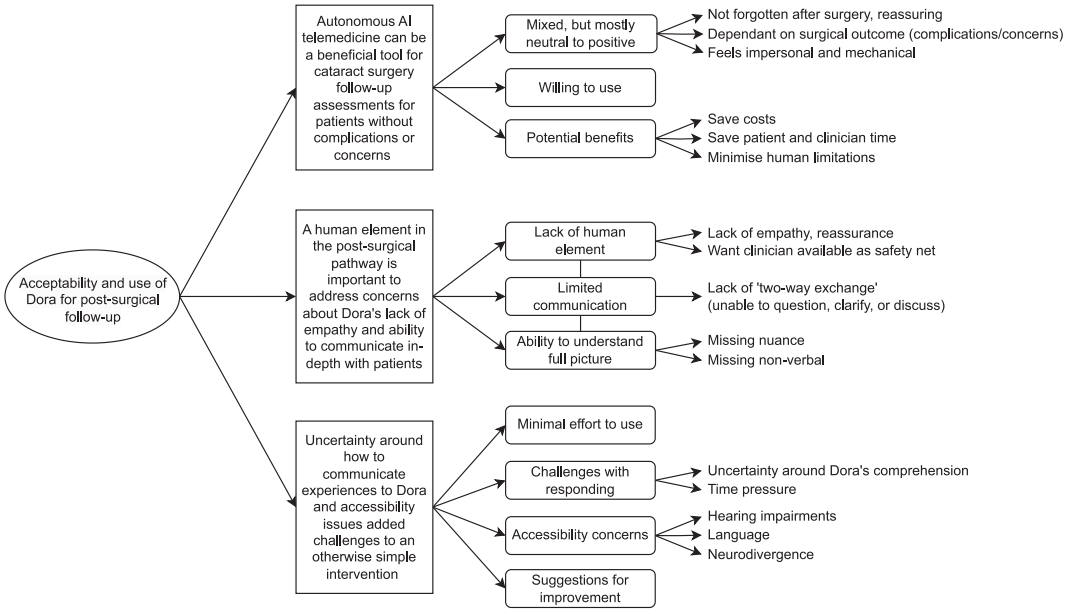

Fig. 5: Thematic map of qualitative interviews to determine acceptability of Dora R1 calls.

by income brackets seven (£60,000–£69,999 per year; rank sum = 5240.5), one (up to £12,499 per year; rank sum = 1661), and two (£12,500–£19,999 per year; rank sum = 1384.5). There were no significant differences ($p < 0.05$) in SUS score by any demographic variable or in any scores by gender or ethnicity (Supplementary

Costs for standard Imperial post-cataract surgery pathway

Total staff cost of face-to-face follow-up	Number of patients	Staff cost/hr	F2F cost per patient	Cost
Nurse consultant (Bands 8a-c Nurse consultant–Average)	35	£83	£41.50	£1452.50
Associate specialist	32	£120	£60	£1920.00
Fellow	12	£52	£26	£312.00
Consultant ophthalmologist	11	£122	£61	£671.00
Specialist trainee (Registrar ST7)	3	£52	£26	£75.66
Optometrists (All)	4	£62	£31	£124.00
Total	97	–	–	£4555.16
Average staff cost per patient	–	–	–	£46.96

Costs for Dora R1 post-cataract surgery pathway (implemented at Imperial)

A. Total cost of face-to-face follow-up	Number of patients	Cost
Nurse consultant (Bands 8a-c Nurse Consultant-Average)	5	£218.33
Associate specialist	3	£200.87
Fellow	1	£24.87
Consultant ophthalmologist	1	£58.35
Specialist trainee (Registrar ST7)	1	£12.43
Optometrists (All)	0	£0.00
Total–A	11	£514.85
B. Total cost of telephone follow-up	Time spent on call in minutes	Cost
Nurse–Band 7	10	£62.00
Total–B	For 55 patients	£568.33
Total staff cost of Dora R1 pathway (A + B)	92	£1084.21
Average staff cost per patient		£11.78

Table 6: **Cost result table for F2F (standard care) versus Dora R1 pathways at Imperial.**

Table S14). Spearman's rank correlation coefficient (Spearman's rho) analysis found no significant linear associations ($p < 0.05$) between age and NPS ($r(185) = 0.21$), SUS ($r(131) = -0.105$), or TUQ ($r(115) = -0.017$) scores. Significant correlations ($p < 0.0001$) were identified between the three usability and acceptability measures (NPS and SUS ($r(123) = 0.379$), NPS and TUQ ($r(108) = 0.585$), SUS and TUQ ($r(110) = 0.520$); Supplementary Figure S4). Bonferroni-adjusted p-values were applied.

Cost analysis

Of the 97 patients at Imperial, 92 patients were called by Dora R1, with 82 calls completed autonomously and 45 patients recommended for review. In a real-world Dora R1 pathway, the 10 incompletes would be booked for callback and were included in cost calculations. Not accounting for the costs of Dora R1 delivery, the analysis found an average staff cost saving of £35.18 per patient (n = 92) compared to standard care (n = 97, Table 6). This includes costs from all of the staff involved in the full clinical pathway in each case.

Discussion

This study examined the potential of the first version of an autonomous natural language agent to deliver routine cataract surgery follow-up assessment calls and detect patients needing further review. Comparing Dora R1 to a supervising clinician provided good preliminary evidence of Dora R1's safety and accuracy through the moderate-to-strong agreement and high accuracy, sensitivity, and specificity. Dora R1 demonstrated potential to improve the efficiency of routine cataract surgery pathways, increase convenience and timeliness for patients, and reduce clinical workload and costs.

Dora R1's decision algorithm was designed to err on the side of caution and provide a high degree of confidence that patients recommended discharge would have no clinical concerns. Only 4 of the 64 patients recommended for review by the supervising clinician were recommended discharge by Dora, none of whom were deemed to require further clinical review on callback. There were 10 patients discharged by both Dora and the clinician who had unexpected management changes within 2 weeks of the Dora R1 call; primarily patients with asymptomatic trace cells detected at planned F2F review at Imperial. It is common not to routinely treat these mild asymptomatic clinical findings.[36–38] There was 1 patient that presented for a delayed unplanned review that had an unexpected management change (rebound acute anterior uveitis at 44 days), however, no patients presented unexpectedly within 2 weeks needing a management change.

One symptom question (vision issues) had a lower sensitivity (Supplementary Table S10); analysing the transcripts found that the binary structure of the question may have made it difficult for participants to articulate their sometimes nuanced symptoms clearly. The algorithm was locked for the duration of the study period, but is one of the features that has been updated in Dora R1's deployment outside the study.

Mixed-methods analyses were conducted to better understand issues that could affect real-world adoption and use of Dora R1. Usability scores were above average and system data demonstrated that most calls were completed autonomously without issue. Although the system took little effort to use, some interviewed participants were unsure how to provide simple responses that captured their experiences. Although several participants expressed a preference for F2F consultations, almost all were willing to use Dora R1 again in routine circumstances. Key concerns revolved around the lack of human element; i.e. that the automated system might not capture all relevant information or be able to deal with patients' emotional reactions.

These findings have implications for Dora R1's adoption and for future applications of AI-enabled telemedicine. In terms of efficiency, Dora R1 successfully recommended discharge for over half of patients (58%); in real-world operation, these patients would not be seen F2F, significantly reducing clinical workload and service delivery costs. For wider implementation, it will be important to communicate to patients the role of Dora R1 and human clinicians within their care and ensure they know how to seek clinician-delivered care if needed.

Limitations included potential issues around selection bias. As patients with strong opposition to AI telemedicine may not have consented to participate, the results could reflect a more positive response than from the general population. Likewise, participants with cognitive difficulties, hearing impairment, or non-English speakers were excluded, reducing generalisability. The two clinical sites included in the study are also located in a similar area in the south of England. Although the Imperial site serves a diverse population, these sites are unlikely to have captured the variety of accents present in the UK, which may influence successful use of Dora R1.

Both sites had COVID-related disruption to elective surgery which reduced anticipated sample sizes, with Imperial having additional disruptions due to a temporary interruption to operating room activity. The disruptions affected the data collection according to the initial health economics analysis plan and a comparison of issues identified during routine follow-up could not be made between the Oxford and Imperial sites. A larger dataset would be required to show the potential cost benefits of implementing Dora R1 and provide further evidence of accuracy and safety. According to the most recent National Ophthalmology Database audit, approximately 2% of cataract surgeries have an intraoperative complication and 5% have a postoperative

complication; Oxford's adjusted posterior capsular rupture (PCR) rate was 0.83% with a case complexity index of 2.50 and Imperial's was 1.27% with a case complexity index of 2.58.[39] Dora R1 is only designed for use for uncomplicated surgeries, but further evaluation in a larger sample could improve the safety and accuracy evidence by ensuring that Dora consistently identifies patients with complications as needing review. Post-operative uveitis and cystoid macular oedema (CMO) are the most common postoperative complications at 3 weeks,[39] which was reflected in our sample. There were no cases of endophthalmitis or retinal detachment at either site during the study period; this is a limitation because these are serious complications of cataract surgery.[39] For Dora to be implemented in clinical practice, it is important to have high confidence that Dora will detect these complications. Another limitation is that Dora R1's decision was compared with only one ophthalmologist's decision. While this is the standard pathway for post-surgical follow-up assessment, the 'ground truth' comparator could have been strengthened by incorporating multiple ophthalmologists' decisions.[40]

The findings provide preliminary evidence that Dora R1 is safe, generally acceptable, and a potentially cost-effective means of delivering cataract surgery follow-up assessments. Larger pragmatic trials should investigate time, cost and CO2 emission-saving benefits (including additional variables such as travel costs and indirect costs associated with time spent attending appointments), barriers and facilitators to implementation, and potential applications of similar technologies to other clinical pathways within and beyond ophthalmology.[41,42] A service evaluation of clinical sites in South East England using Dora R3 at various points in the cataract pathway (pre- and post-surgery) is underway and expected to provide evidence around implementation, operational efficiency, and patient outcomes.[43]

Surgery outcome was a key factor related to patients' attitudes towards Dora R1. Although no interviewees experienced severe complications, several felt they would be less comfortable using Dora R1 if they had. It will be important to ensure that introducing automation into healthcare does not dehumanise care by removing patient–clinician interaction, but rather increases clinicians' interactions with patients who have concerns.[44,45] Studies implementing similar systems in other contexts should consider the likelihood of complications and patients' post-surgical emotional experience; mitigations such as more patient information, a clear procedure for what to do if concerned, and tailored call timelines might be necessary for acceptability.

The study also highlighted the importance of digital equity and accessibility. Issues of comprehension and hearing are common for natural language technologies and implementation will require equitable alternatives for non-English speakers and people with serious hearing difficulties.[46] Another digital equity issue that arose was related to the acceptability of this technology for people with autism; although Dora R1 was generally acceptable, it was very unacceptable for a participant with autism. This case raises two key points: potential issues of acceptability for neurodiverse individuals using automated conversational agents, which should be investigated further and accounted for in implementation, and the issue of inclusivity in research. If the participant with autism had not reached out to the research team to provide feedback, they likely would have been missed from the randomly-selected subset of participants for interview. Deliberate efforts are needed to include diverse perspectives through the digital health development and evaluation process[47] to ensure that different needs are understood and accounted for. For instance, digital equity and accessibility issues could be facilitated by including a brief question in pre-surgical assessments to determine whether patients are willing and able to receive the automated call and by delivering digital services in multiple languages.

In conclusion, our findings provide preliminary evidence of Dora R1's clinical safety and potential for adoption. Although patients expressed some concerns about the lack of human involvement and the more limited scope of Dora R1's assessment, the evidence supports Dora R1 as a generally acceptable, safe, and potentially cost-beneficial alternative to clinician-delivered cataract surgery follow-up assessments. Further research should evaluate Dora R1's effectiveness and monitor safety in the real-world, unsupervised context in which it would be implemented. More broadly, our findings highlight the potential for automated systems to support other repetitive, low-skill healthcare tasks and reduce burden on healthcare professionals. We identified key factors that could affect their successful adoption, including accessibility and a need for a human element in care, particularly for patients experiencing complications. These should be considered when implementing the intervention - in cataract surgery and other contexts–to ensure that all patients receive equitable, humanising care.

Contributors

NdP conceived of the study topic. The study design was developed by NdP, GM, EL, MMI, EN, KX, EM and AH. KX and EN led patient recruitment at the two study sites. AH and EL supervised Dora R1's calls and compiled and verified the clinical data. MMI conducted the usability survey and MMI and EM completed participant interviews. Statistical analysis of the quantitative data was conducted by the University of Plymouth, MMI, SB and EM conducted the thematic analysis of the qualitative data. MMI and EM accessed and verified the usability and qualitative data. MB was responsible for the cost analysis. EM set the structure for the manuscript, MMI drafted based on this framework, and it was reviewed by all authors with final revisions from all authors.

Data sharing statement

The datasets generated and analysed during the current study are not publicly available due to them containing information that could compromise research participant privacy/consent. Deidentified individual participant data, data dictionaries, and other study materials will

be available upon publication for 10 years from the corresponding authors [EN, KX] on reasonable request.

Declaration of interests

At the time of the study NdP and GM were employees of Ufonia Limited, a voice artificial intelligence company, NdP as Director and Shareholder. GM was also an employee of Oxford University Hospitals during the period of research; he is no longer an employee of Ufonia Limited and holds no shares. Since the completion of the study AH and EL have become employees of Ufonia Limited. During part of the study period NdP was employed part-time by Oxford University Hospitals NHS Foundation Trust and responsible for the development of innovation activities at the Trust. His work for Ufonia was approved by the Trust and was declared in the Trust's register of interests. None of the resources he was responsible for within the Trust were used to support the project. During part of the study period he was also Clinical Lead for the Thames Valley & Surrey Local Health and Care Records Programme, his work for Ufonia has been declared in the Programme's register of interests. EM, MMI, SB, MB, EN and KX declare no financial or non-financial conflicts of interest.

Acknowledgements

This manuscript is independent research funded by the National Institute for Health Research (NIHR) and NHSX (Artificial Intelligence in Health and Care Award, AI_AWARD01852). KX is funded by the Wellcome Trust (216593/Z/19/Z) and NIHR Oxford Biomedical Research Centre (BRC). EM, MMI, and SB are supported by the NIHR Newcastle BRC based at the Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle University, and the Cumbria, Northumberland, and Tyne and Wear NHS Foundation Trust. The open access publication fee was paid from the Imperial College London Open Access Fund. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, Wellcome Trust, NHS, Department of Health and Social Care, Ufonia Limited, or any of the authors' affiliated universities or BRCs. The funding body was not involved in the study design, data collection or analysis, or the writing and decision to submit the article for publication.

Ufonia Limited has been supported by grant funding from Innovate UK (part of UK Research and Innovation), the Science and Technology Facilities Council and the National Institute for Health and Care Research (NIHR). The company has received business support from Oxford University Innovation and the Oxford Foundry. The Department of Ophthalmology at Buckinghamshire Healthcare NHS Trust has collaborated in the development of Ufonia's system. We would like to acknowledge great efforts by the statistical analysis team at Precision Consulting and the clinical trial units at both Oxford (Eye Research Group Oxford, ERGO) and Imperial College in participant recruitment and data collection over the challenging COVID pandemic period. In particular, from Oxford: Alexina Fantato (0009-0009-8659-2679), Caroline Justice, and Janette Savage and from Imperial: Faisal Ahmed (0000-0003-0015-6008), Zena Rodrigues (0000-0003-1701-4129), Jessica Bonetti (0000-0002-0297-2132), Serge Miodragovic (0000-0002-8632-7383), Ali Mearza (0000-0002-0488-6107), Imran Mohammed (0009-0001-4503-2829), Giuseppe Serrano (0009-0005-4411-4194), and Matt Mahey.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102692.

References

1 *NHS long term workforce plan*. NHS England; 2023. https://www.england.nhs.uk/long-read/nhs-long-term-workforce-plan-2/#references. Accessed May 15, 2024.
2 Department of Health and Social Care. *Digital revolution to bust COVID backlogs and deliver more tailored care for patients*. GOVUK; 2022. https://www.gov.uk/government/news/digital-revolution-to-bust-covid-backlogs-and-deliver-more-tailored-care-for-patients. Accessed August 16, 2022.
3 *22/26 HSDR evaluating the high volume low complexity (HVLC) surgical hubs model commissioning brief*. National Institute for Health and Care Research; 2022. https://www.nihr.ac.uk/documents/2226-hsdr-evaluating-the-high-volume-low-complexity-hvlc-surgical-hubs-model-commissioning-brief/29940. Accessed August 16, 2022.
4 Oliver D, Foot C, Humphries R. *Making our health and care systems fit for an ageing population*. The King's Fund; 2014.
5 Myers LC, Liu VX. The COVID-19 pandemic strikes again and again and again. *JAMA Netw Open*. 2022;5:e221760. https://doi.org/10.1001/jamanetworkopen.2022.1760.
6 Alderwick H. Is the NHS overwhelmed? *BMJ*. 2022;376. https://doi.org/10.1136/bmj.o51.
7 Dall TM, Gallo PD, Chakrabarti R, West T, Semilla AP, Storm MV. An aging population and growing disease burden will require ALarge and specialized health care workforce by 2025. *Health Aff*. 2017;32(11):2013–2020. https://doi.org/10.1377/hlthaff.2013.0714.
8 Donachie PHJ, Sparrow JM. *National ophthalmology database audit: year 5 annual report*. The Royal College of Ophthalmologists; 2020.
9 [No title] n.d https://www.rcophth.ac.uk/wp-content/uploads/2016/07/Managing-an-outbreak-of-postoperative-endophthalmitis.pdf. Accessed November 3, 2020.
10 Aaronson A, Viljanen A, Kanclerz P, Grzybowski A, Tuuminen R. Cataract complications study: an analysis of adverse effects among 14,520 eyes in relation to surgical experience. *Ann Transl Med*. 2020;8. https://doi.org/10.21037/atm-20-845.
11 Haripriya A, Chang DF, Reena M, Shekhar M. Complication rates of phacoemulsification and manual small-incision cataract surgery at Aravind Eye Hospital. *J Cataract Refract Surg*. 2012;38. https://doi.org/10.1016/j.jcrs.2012.04.025.
12 de Pennington N, Mole G, Lim E, et al. Safety and acceptability of a natural language artificial intelligence assistant to deliver clinical follow-up to cataract surgery patients: proposal. *JMIR Res Protoc*. 2021;10. https://doi.org/10.2196/27227.
13 *The way forward: executive summary*. The Royal College of Ophthalmologists; 2017.
14 Moustafa GA, Borkar DS, Borboli-Gerogiannis S, et al. Optimization of cataract surgery follow-up: a standard set of questions can predict unexpected management changes at postoperative week one. *PLoS One*. 2019;14:e0221243. https://doi.org/10.1371/journal.pone.0221243.
15 Khavandi S, Lim E, Mole G. 110 Patient acceptability of telephone follow up after cataract surgery. *BMJ Leader*. 2020;4. https://doi.org/10.1136/leader-2020-FMLM.110.
16 Khavandi S, Lim E, Higham A, et al. User-acceptability of an automated telephone call for post-operative follow-up after uncomplicated cataract surgery. *Eye*. 2022;37(10):2069–2076. https://doi.org/10.1038/s41433-022-02289-8.
17 Hamilton AB, Mittman BS. *Implementation science in health care*. Oxford University Press; 2017. https://doi.org/10.1093/oso/9780190683214.003.0023.
18 Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: decide-ai. *Nat Med*. 2022;28:924–933. https://doi.org/10.1038/s41591-022-01772-9.
19 O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. *Acad Med*. 2014;89:1245–1251. https://doi.org/10.1097/ACM.0000000000000388.
20 Blank G. *OxIS 2019 questionnaire*. All parts; 2019.
21 Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. *BMC Health Serv Res*. 2017;17:1–13. https://doi.org/10.1186/s12913-017-2031-8.
22 *College Station, TX: StataCorp LLC*. Stata Statistical Software; 2021.
23 Lewis JR. The system usability scale: past, present, and future. *Int J Hum Comput Interact*. 2018;34:577–590. https://doi.org/10.1080/10447318.2018.1455307.
24 Parmanto B, Lewis AN JR, Graham KM, Bertolet MH. Development of the Telehealth usability questionnaire (TUQ). *Int J Telerehabilitation*. 2016;8:3. https://doi.org/10.5195/ijt.2016.6196.
25 What is net promoter? Net promoter network. n.d https://www.netpromoter.com/know/. Accessed August 16, 2022.
26 Brooks J, McCluskey S, Turley E, King N. The utility of template analysis in qualitative psychology research. *Qual Res Psychol*. 2015;12:202–222. https://doi.org/10.1080/14780887.2014.955224.
27 Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher. *Int J Transgend Health*. 2022;24(1):1–6. https://doi.org/10.1080/26895269.2022.2129597.
28 NVivo (released in March 2020). *QSR international pty Ltd*. 2020.

29 Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3:77–101. https://doi.org/10.1191/1478088706qp063oa.

30 Saldaña J. *The coding manual for qualitative researchers*. SAGE; 2021.

31 Jones KC, Burns A. *Unit costs of health and social care 2021*. Personal Social Services Research Unit; 2021. https://doi.org/10.22024/UNIKENT/01.02.92342 [object Object].

32 Borkar DS, Laíns I, Eton EA, et al. Incidence of management changes at the postoperative week 1 visit after cataract surgery: results from the perioperative care for IntraOcular lens study. *Am J Ophthalmol*. 2019;199:94–100. https://doi.org/10.1016/j.ajo.2018.10.013.

33 McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22:276.

34 Baratloo A, Hosseini M, Negida A, El Ashal G. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. *Emergency*. 2015;3:48.

35 *What is a good net promoter score?* Qualtrics; 2021. https://www.qualtrics.com/uk/experience-management/customer/good-net-promoter-score/. Accessed August 19, 2022.

36 Reddy AK, Patnaik JL, Miller DC, Lynch AM, Palestine AG, Pantcheva MB. Risk factors associated with persistent anterior uveitis after cataract surgery. *Am J Ophthalmol*. 2019;206:82–86. https://doi.org/10.1016/j.ajo.2019.02.016.

37 Neatrour K, McAlpine A, Owens TB, Trivedi RH, Poole Perry LJ. Evaluation of the etiology of persistent iritis after cataract surgery. *J Ophthalmic Inflamm Infect*. 2019;9:4. https://doi.org/10.1186/s12348-019-0170-2.

38 Miller KM, Oetting TA, Tweeten JP, et al. Cataract in the adult eye preferred practice pattern. *Ophthalmology*. 2022;129:P1–P126. https://doi.org/10.1016/j.ophtha.2021.10.006.

39 Donachie PHJ, Buchan JC. *National ophthalmology database audit: year 7 annual report – the sixth prospective report of the national ophthalmology database audit national cataract audit*. The Royal College of Ophthalmologists; 2023.

40 Chen P-HC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *Lancet Digital Health*. 2021;3:e693–e695. https://doi.org/10.1016/S2589-7500(21)00216-8.

41 Best practice library – pathways – getting it right first time – GIRFT n.d. https://www.gettingitrightfirsttime.co.uk/bpl/pathways/. Accessed September 23, 2022.

42 Betzler BK, Chen H, Cheng C-Y, et al. Large language models and their impact in ophthalmology. *Lancet Digital Health*. 2023;5:e917–e924. https://doi.org/10.1016/S2589-7500(23)00201-7.

43 Automated cataract surgery telephone consult evaluation collaboration, Higham A. Automated clinical conversations across the cataract pathway with an artificial intelligence (AI) conversation agent: a UK regional service evaluation protocol. *bioRxiv*. 2023. https://doi.org/10.1101/2023.06.14.23291399.

44 Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. *NPJ Digital Med*. 2018;1:1–4. https://doi.org/10.1038/s41746-017-0012-2.

45 D'Amario D, Canonico F, Rodolico D, et al. Telemedicine, artificial intelligence and humanisation of clinical pathways in heart failure management: back to the future and beyond. *Card Fail Rev*. 2020;6. https://doi.org/10.15420/cfr.2019.17.

46 Morris MR. *AI and accessibility*; 2020. https://cacm.acm.org/magazines/2020/6/245157-ai-and-accessibility/fulltext. Accessed September 23, 2022.

47 Iakovleva T, Oftedal E, Bessant J. Changing role of users—innovating responsibly in digital health. *Sustain Sci Pract Policy*. 2021;13:1616. https://doi.org/10.3390/su13041616.