A machine learning system to optimise triage in an adult ophthalmic emergency department: a model development and validation study

Camilo Brandao-de-Resende,[a,b,c,*] Mariane Melo,[b,c] Elsa Lee,[a,b,c] Anish Jindal,[a,d] Yan N. Neo,[d] Priyanka Sanghi,[d] Joao R. Freitas,[c,e] Paulo V. I. P. Castro,[f] Victor O. M. Rosa,[f] Guilherme F. S. Valentim,[f] Maria Luisa O. Higino,[f] Gordon R. Hay,[a,d] Pearse A. Keane,[a,g] Daniel V. Vasconcelos-Santos,[f] and Alexander C. Day[a,d]

[a]Institute of Ophthalmology, University College London (UCL), London, UK
[b]NIHR Moorfields Clinical Research Facility, Moorfields Eye Hospital NHS Foundation Trust, London, UK
[c]Research Department, DemDX Ltd, London, UK
[d]Accident and Emergency Department, Moorfields Eye Hospital NHS Foundation Trust, London, UK
[e]University of Sao Paulo (USP), Sao Paulo, Brazil
[f]Hospital Sao Geraldo, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil
[g]NIHR Moorfields Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK

eClinicalMedicine
2023;66: 102331
Published Online 2 December 2023
https://doi.org/10.1016/j.eclinm.2023.102331

Summary

Background A substantial proportion of attendances to ophthalmic emergency departments are for non-urgent presentations. We developed and evaluated a machine learning system (DemDx Ophthalmology Triage System: DOTS) to optimise triage, with the aim of reducing inappropriate emergency attendances and streamlining case referral when necessary.

Methods DOTS was built using retrospective tabular data from 11,315 attendances between July 1st, 2021, to June 15th, 2022 at Moorfields Eye Hospital Emergency Department (MEH) in London, UK. Demographic and clinical features were used as inputs and a triage recommendation was given (“see immediately”, “see within a week”, or “see electively”). DOTS was validated temporally and compared with triage nurses’ performance (1269 attendances at MEH) and validated externally (761 attendances at the Federal University of Minas Gerais - UFMG, Brazil). It was also tested for biases and robustness to variations in disease incidences. All attendances from patients aged at least 18 years with at least one confirmed diagnosis were included in the study.

Findings For identifying ophthalmic emergency attendances, on temporal validation, DOTS had a sensitivity of 94.5% [95% CI 92.3–96.1] and a specificity of 42.4% [38.8–46.1]. For comparison within the same dataset, triage nurses had a sensitivity of 96.4% [94.5–97.7] and a specificity of 25.1% [22.0–28.5]. On external validation at UFMG, DOTS had a sensitivity of 95.2% [92.5–97.0] and a specificity of 32.2% [27.4–37.0]. In simulated scenarios with varying disease incidences, the sensitivity was ≥92.2% and the specificity was ≥36.8%. No differences in sensitivity were found in subgroups of index of multiple deprivation, but the specificity was higher for Q2 when compared to Q4 (Q4 is less deprived than Q2).

Interpretation At MEH, DOTS had similar sensitivity to triage nurses in determining attendance priority; however, with a specificity of 17.3% higher, DOTS resulted in lower rates of patients triaged to be seen immediately at emergency. DOTS showed consistent performance in temporal and external validation, in social-demographic subgroups and was robust to varying relative disease incidences. Further trials are necessary to validate these findings. This system will be prospectively evaluated, considering human-computer interaction, in a clinical trial.

Funding The Artificial Intelligence in Health and Care Award (AI_AWARD01671) of the NHS AI Lab under National Institute for Health and Care Research (NIHR) and the Accelerated Access Collaborative (AAC).

Keywords: Ophthalmology; Triage; Machine learning; Artificial intelligence; Emergency care; Digital health

*Corresponding author. Institute of Ophthalmology, University College London (UCL), London EC1V 9EL, UK.
E-mail address: camilo.deresende@nhs.net (C. Brandao-de-Resende).

Research in context

Evidence before this study

We searched PubMed/MEDLINE for relevant work published between Jan 1, 1980, and Oct 3, 2023, with no language restrictions, using the terms (“triage” [MeSH Terms] OR “triage” [Title/Abstract]) AND (“ophthalmology” [MeSH Terms] OR “ophthalmology” [Title/Abstract] OR “eye” [Title/Abstract]) AND (“machine learning" [MeSH Terms] OR “machine learning" [Title/Abstract] OR “deep learning" [Title/Abstract] OR “artificial intelligence" [MeSH Terms] OR “artificial intelligence" [Title/Abstract] OR “algorithm" [Title/Abstract]). This search yielded 39 studies, most of which were reviews or related with triaging a single disease (such as diabetic retinopathy). The majority of the studies evaluating automated triage were small, lacked relevant metrics (such as specificity and sensitivity), and did not present temporal or external validation or fairness (or bias or trustworthiness) analyses. One study proposed a self-triage model using metadata and smartphone images but was tested only on 103 patients, included only 18 possible differentials, and did not consider the potential increase of non-urgent presentations to emergency departments, aggravating professional burden and increasing healthcare costs.

Added value of this study

In this study, we developed and evaluated a novel machine learning system to optimise ophthalmic triage, using data from 12,494 patients with 95 differentials. The system was built to be used by triage nurses, without disrupting their workflow, minimising potential harms caused by potential failures, and optimising resource utilisation. Our tabular dataset used to build the system consists of presentations to Europe’s largest eye emergency department (in London, UK) over a 1-year period encompassing anterior and posterior segment pathologies. The system was validated temporally and externally (in a different country) using relevant metrics (e.g. sensitivity and specificity), compared with triage nurses performance, and tested for biases and robustness to variations in disease incidences. In addition, explainability was assessed and potential serious misses disclosed.

Implications of all the available evidence

The developed system had similar sensitivity to triage nurses in determining attendance priority (~ 95%); however, with a specificity up to 17.3% higher, the model resulted in lower rates of patients triaged to be seen immediately at emergency. It showed consistent performance levels in different countries and in socio-demographic subgroups, being fair, and robust to varying disease incidences in simulated scenarios. Potentially, with further validation, it could reduce the number of non-urgent patients who are triaged to be seen on the same day; thus reducing the emergency costs in a safe way. The model was created to be compatible with clinical workflows. Further trials are necessary. This system will be prospectively evaluated, considering human-computer interaction, in a clinical trial.

Introduction

Ophthalmology is the busiest speciality by outpatient workload in the United Kingdom (UK) and demand for eye care has outgrown the workforce worldwide.[1,2] New attendances in emergency departments have been on the rise, accounting for over 30% of all ophthalmic attendances in National Health Service (NHS) England.[3,4] Notably, the proportion of patients presenting with non-urgent conditions is significant (>15%) yet providing non-urgent care can be three times more costly than similar attendances in other settings.[5,6]

Machine learning (ML) is being increasingly applied to improve clinical workflow efficiency and has the potential to enhance the accuracy of triage, optimising service allocation.[7] Within triage, ML has the capability to process high dimensionality structured data and the potential to achieve superior performance compared to rule-based algorithms by abstracting complex non-linear patterns between patients’ clinical presentation and their clinical risk. One study proposed an ophthalmic self-triage model using metadata and smartphone images but was tested only on 103 patients, included only 18 possible differentials, and did not consider the potential increase of non-urgent presentations to emergency departments, aggravating professional burden and increasing healthcare costs.[8]

The purpose of this study was to develop and evaluate the performance, fairness, and robustness of a ML-based system to optimise triage in ophthalmic emergency department. The support triage platform is called DemDx Ophthalmology Triage System (DOTS). It aims to help nurses to make more accurate and safe triage decisions, thus reducing the proportion of cases unnecessarily requiring same day review and streamlining non-urgent ones to other settings for appropriate management in a more efficient way.

Methods

Ethics

The study was approved by the Institutional Review Boards at Moorfields Eye Hospital NHS Trust (MEH), London, UK, (IRAS ID: 290843) and at the Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil (CAAE 49591321.9.0000.5149). DOTS was developed and tested following the principles of the

Declaration of Helsinki and the Ethics Guidelines for Trustworthy Artificial Intelligence from the European Commission. No informed consent was required as only deidentified data was used. The study follows the reporting standards set out in the TRIPOD guideline for Prediction Model Development and Validation.[9]

Data

For development and temporal validation, deidentified data were collected at the emergency department of MEH, from July 1st, 2021, to June 15th, 2022. Triage nurses used automated forms to input patient data, including laterality, duration, red flags, signs/symptoms, ocular/systemic comorbidities, and outcome (details in Supplementary material). Specific data inputs were laterality (unilateral, bilateral), duration (<24 h, 2–3 days, 4–7 days, 1–4 weeks, >4 weeks), red flags (rapid change in visual acuity, complete visual loss, diplopia, change in pupils, systemic unwellness, post-op, or no red flags), signs and symptoms, history, and triage nurse outcome (see in emergency, walk to speciality clinics, treatment/advice given at triage only, referred to Urgent Care Clinic (UCC) for review within a week, to see General Practitioner (GP), or see Optometrist).

Data collection form was developed in three main steps: (1) Prototype based on MEH current form, (2) creation of list of signs and symptoms based on ophthalmic symptoms present on BMJ Best practices,[10] reviewed by three emergency consultants, two triage nurses, and one optometrist, (3) refinement after feedback from other 10 nurses and three emergency consultants started on July 1st 2021. Revised form was frozen on August 08th 2021.

Demographic data and diagnoses were collected from the electronic medical record (EMR). EMR data included gender, age bracket, ethnicity following the NHS grouping system,[11] the Index of Multiple Deprivation (IMD)[12] (determined by the first three digits of the post code and grouped by quintiles from Q1 to Q5, where Q1 is the most deprived and Q5 is the least deprived). Patients that were seen by an ophthalmologist had their diagnoses extracted from the EMR. Patients discharged by triage nurses without seeing an ophthalmologist, had their cases reviewed by two authors, who tried to determine a diagnosis based on history and referral letters.

Exclusion criteria included (1) data collected before August 08th 2021 (during data collection form refinement), (2) incompatibility of input from data collection form and the EMR (representing input error), (3) patients <18 years old, and (4) attendances with no diagnosis in the EMR or after review (note: no abnormality detected is a possible diagnosis).

Diagnoses were classified as elective, urgency (see within a week), or emergency (see immediately), accordingly to consensus among three Accident and Emergency consultants (see Supplementary material). The priority for each attendance (ground-truth) was determined accordingly to the diagnosis with highest priority present in the attendance.

For nurses, priority labels were determined based on the triage outcome. “See in Emergency” and “Walk to speciality clinics” were classified as emergency. “See in UCC” was classified as urgency. “Treated/advice given at triage only”, “To see GP”, and “To see Optometrist” were classified as elective.

Attendances from new patients after May 1st, 2022, were included in the temporal validation (TVal) dataset. Attendances between August 09th 2021 and April 30th, 2022, were randomly split into training (85%) and internal validation (Val) (15%) datasets. Attendances from a single patient were always included in the same dataset.

For external validation (Eval), deidentified data from consecutive patients were collected from the EMR of UFMG ophthalmic emergency department, from April 20th, 2023, to May 11th, 2023. This is a public university-based referral centre, receiving patients not only from Belo Horizonte, the capital of the state of Minas Gerais (MG), Southeastern Brazil, but from most of the 853 cities in the state. MG recapitulates socioeconomical diversity seen on a national scale. Records from triage nurses were analysed by five doctors, who structured the clinical features described and translated to the equivalent ones in DOTS (Supplementary material). End differentials were collected from medical records. Since at UFMG the triage nurses only define the order of attendances, all patients were seen by doctors at the same day. Inclusion and exclusion criteria were the same as for MEH data.

All data were deidentified before being transferred to a secure server on an AWS cloud.[13] Details of data collection and data preparation are shown in Supplementary material.

System framework

Based on demographics and clinical features (inputs), a triage recommendation is given based on predicted priority from a multi-class classification model, as shown in Fig. 1. The system outputs a 3-class priority recommendation: emergency (immediately), urgency (see within a week), or elective (Advise/See GP/See Optometrist).

Development of the model

All models were developed using Python 3.8.10.[14] Four architectures were evaluated, chosen to include simple ones as baseline (Logistic Regression, LogisticRegression from sklearn 0.24.1, and Decision Tree, DecisionTreeClassifier from sklearn 0.24.1) as well as the ones considered the state-of-the-art for tabular data, including tree bagging (Random Forest, RandomForestClassifier from sklearn 0.24.1) and tree boosting (XGBoost, xgboost 1.5.2).[15] Details about the hyperparameter selection and tuning of each model can be found in Table 6.2 of the Supplementary material.

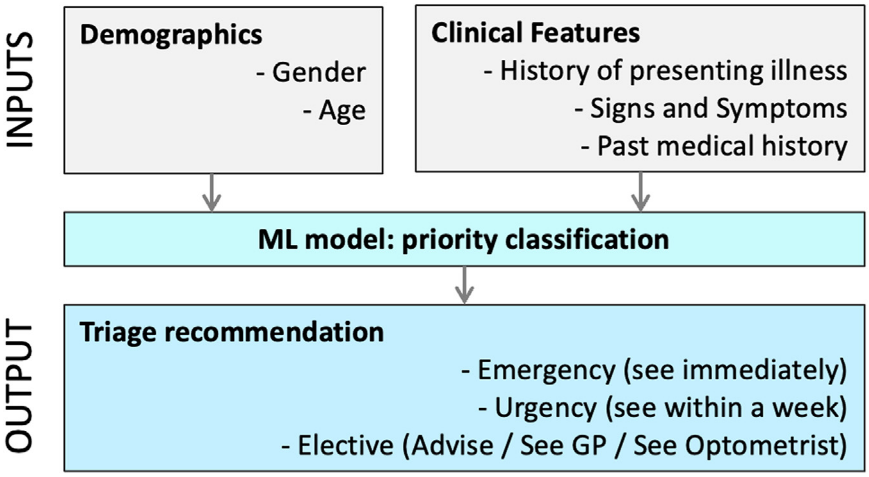
INPUTS
Demographics
- Gender
- Age
Clinical Features
- History of presenting illness
- Signs and Symptoms
- Past medical history
ML model: priority classification
OUTPUT
Triage recommendation
- Emergency (see immediately)
- Urgency (see within a week)
- Elective (Advise / See GP / See Optometrist)

Fig. 1: Architecture of the proposed machine learning-based ophthalmic triage system. Demographics and clinical features form the inputs. The model outputs a 3-class priority recommendation. GP: general practitioner.

We present here the results for emergency predictions. Any elective or urgent case predicted as emergency were considered false positives while any emergency predicted as urgency or elective were considered false negatives. The analysis for urgencies can be found in Supplementary material.

DOTS was developed and validated for real-world application, where we cannot rely on areas under a receiver operating characteristic (ROC) curve to deploy a model, and therefore thresholds for probabilities should be pre-defined. These thresholds determine the specific points over the ROC curves at which the deployed model is operating. Even ROC curves with similar areas might have drastically different real-world performance if the thresholds shift under different settings. Therefore, we defined thresholds for decisions during the development phase, subsequently used these in all evaluations, and compared relevant metrics for triage systems (sensitivity, specificity, PPV, and NPV).

In hyperparameter selection, probability thresholds for emergency and urgency were sequentially determined, so that a sensitivity ≥ the lower boundary of the 95% confidence interval (CI) of the nurse sensitivity was guaranteed on the Val dataset. Specificities using the thresholds were then calculated. The model with the highest weighted average of specificities for emergency and urgency on Val dataset was selected as best-performing.

Temporal and external validation

DOTS was tested on TVal and EVal datasets. It was evaluated using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and positive rate. Same metrics were estimated for the triage nurses from MEH using the reported outcomes on TVal.

Explainability, fairness and trustworthiness

DOTS was explained using SHapley Additive exPlanations (SHAP).[16] Fairness was evaluated in demographic subgroups, including gender, age, ethnicity, and IMD. Besides the external validation, robustness to changes in the relative disease incidences was evaluated by comparing six simulated scenarios built from subsamples of TVal dataset.

Statistical analysis

Assuming an expected sensitivity of 95% for emergencies and that 45% of the cases are true emergencies, a sample size of 1156 observations is sufficient to achieve a significance of 5% with an acceptable error of 2% (see Supplementary material). The training, Val, and TVal datasets were designed to be larger than the estimated sample size.

Statistical analysis was done using Python 3.8.10.[14] A significance level of 5% was considered. All variables were binary and described as number (%). The 95% CI were described as [lower boundary, upper boundary]. Two-sided Fisher's exact tests (fisher_exact from scipy 1.6.2) were used to compare proportions and CIs were determined by binomial proportion (proportion_confint from stats models 0.12.2).

Role of the funding source

The funders were not involved in the study design; the collection, analysis, interpretation, and reporting of data; the writing of the manuscript; or the decision to submit for publication.

Results

Data description

A total of 12,584 attendances from 11,733 patients from MEH were included in the study (training, internal validation [Val], and temporal validation [TVal] datasets). Twenty-four triage nurses collaborated on the study; the number triaged per nurse varied from 156 to 1088 (average 524). The training dataset included 9850 (78.3%) attendances from 9045 (77.1%) patients, Val included 1465 (11.6%) from 1459 (12.4%) patients, and TVal included 1269 (10.1%) from 1229 (10.5%) patients. 209 (2.1%) re-attendances occurred within 2 weeks of initial presentation.

The external validation (Eval) dataset included a total of 761 attendances from 761 patients from UFMG, Brazil. Demographic characteristics are shown in Table 1.

At MEH, a total of 14,996 diagnoses were made (average 1.15/attendance). The most common diagnoses were dry eyes (10.2%), conjunctival/corneal injury (10.2%), and blepharitis (8.0%). Nurses decided that 10,678 (84.9%) attendances should be seen same day in emergency, within those, the most common diagnoses were conjunctival/corneal injury (11.8%), dry eyes (9.3%), anterior uveitis (8.7%), keratitis (7.2%), and blepharitis (6.8%). Among 553 (4.4%) attendances triaged to UCC, the most common diagnoses were dry eyes (20.3%), blepharitis (18.8%), cataract or posterior capsular opacification (10.1%), no abnormality detected (5.5%), and posterior vitreous detachment (5.3%).

At UFMG, a total of 937 diagnoses were made (average of 1.10 per attendance). The most common diagnoses were conjunctival/corneal injury (27.8%), infective conjunctivitis (16.0%), and blepharitis (8.1%).

Priority labels were based on the diagnoses. At MEH, 5731 (45.5%) attendances were labelled as emergency, 1416 (11.3%) as urgency, and 5437 (43.2%) as elective. MEH nurse triage outcomes by label (actual priority) are shown in Table 2. At UFMG, 392 (51.5%) attendances were labelled as emergency, 22 (2.9%) as urgency, and 347 (45.6%) as elective. Complete data description is available in the Supplementary material.

Performance and validation

The XGboost model was selected as the one with the best performance on Val dataset (see Supplementary material). The performance of DOTS compared to triage nurses from MEH for emergency cases, using pre-defined thresholds, is shown in Table 3.

Fig. 2 shows the ROC curves for DOTs in different datasets. For comparison, on TVal, changing DOTS operating point to match the sensitivity of triage nurses (96.4%), it presented a specificity of 32.6% [95% CI 29.2–36.2], compared to 25.1% [22.0–28.5] for triage nurses. More details on model performance are shown in Supplementary material.

Explainability

The most important inputs for the classification of different priorities by DOTS, determined using the Val dataset, are shown in Fig. 3.

Robustness to changes in disease incidences

In addition to the external validation, DOTS robustness to changes in the relative disease incidences was evaluated in simulated scenarios obtained from the original TVal (baseline). The following scenarios were included, increasing the relative proportion of different

	Level	MEH Data (training, Val, TVal) (N = 12,584)	UFMG Data (EVal) (N = 761)
Gender	Female	6620 (52.6)	322 (42.3)
	Male	5964 (47.4)	439 (57.7)
Age (years)	[18,30]	2210 (17.6)	132 (17.3)
	[30,50]	4638 (36.9)	272 (35.7)
	[50,70]	4341 (34.5)	284 (37.3)
	(70,+)	1395 (11.1)	73 (9.6)
Ethnicity	Asian	1193 (9.5)	Not available
	Black	1009 (8.0)	
	Mixed	72 (0.6)	
	Unknown	2168 (17.2)	
	Other	5197 (41.3)	
	White	2945 (23.4)	
Index of multiple deprivation (IMD) Quintile (Q1 is the most deprived and Q5 is the least deprived)	Q1	739 (5.9)	Not available
	Q2	6013 (47.8)	
	Q3	2928 (23.3)	
	Q4	2122 (16.9)	
	Q5	669 (5.3)	
	Unknown	113 (0.9)	

MEH: Moorfields Eye Hospital NHS Trust, London, UK; TVal: temporal validation; UFMG: Federal University of Minas Gerais, Belo Horizonte, Brazil; EVal: external validation.

Table 1: **Demographics of all included attendances.**

Priority label (Ground-truth)	Nurse triage outcome	Total (N = 12,584)	Train/Val (N = 11,315)	TVal (N = 1269)
Emergency	Total	5731 (45.5)	5170 (45.7)	561 (44.2)
	See in emergency	5497 (95.9)	4969 (96.1)	528 (94.1)
	Treated/advice given at triage only	81 (1.4)	72 (1.4)	9 (1.6)
	See in UCC	66 (1.2)	55 (1.1)	11 (2.0)
	Walk to speciality clinics	48 (0.8)	35 (0.7)	13 (2.3)
	To see GP	35 (0.6)	35 (0.7)	0 (0.0)
	To see optometrist	4 (0.1)	4 (0.1)	0 (0.0)
Urgency	Total	1416 (11.3)	1237 (10.9)	179 (14.1)
	See in emergency	1276 (90.1)	1122 (90.7)	154 (86.0)
	See in UCC	73 (5.2)	59 (4.8)	14 (7.8)
	Treated/advice given at triage only	41 (2.9)	36 (2.9)	5 (2.8)
	Walk to speciality clinics	12 (0.8)	7 (0.6)	5 (2.8)
	To see GP	11 (0.8)	10 (0.8)	1 (0.6)
	To see optometrist	3 (0.2)	3 (0.2)	0 (0.0)
Elective	Total	5437 (43.2)	4908 (43.4)	529 (41.7)
	See in emergency	3905 (71.8)	3539 (72.1)	366 (69.2)
	Treated/advice given at triage only	771 (14.2)	665 (13.5)	106 (20.0)
	See in UCC	414 (7.6)	387 (7.9)	27 (5.1)
	To see GP	274 (5.0)	252 (5.1)	22 (4.2)
	To see optometrist	43 (0.8)	40 (0.8)	3 (0.6)
	Walk to speciality clinics	30 (0.6)	25 (0.5)	5 (0.9)

TVal: temporal validation; UCC: Urgent Care Centre; GP: General Practitioner.

Table 2: **Priority label defined by diagnoses vs nurse triage outcome.**

conditions: (1) Inflammatory conditions, (2) Vitreous and retinal detachment, (3) Trauma and keratitis, (4) Conjunctivitis, (5) Elective conditions.

No statistical differences existed in sensitivity or specificity for emergency or urgency differentials between any scenario and the original TVal. Details and full results including urgencies are shown in the Supplementary material.

Fairness

The fairness of DOTS was evaluated by analysing the sensitivity and specificity in demographic subgroups of TVal (gender, age, ethnicity, and IMD) and EVal (gender and age).

On TVal, for emergencies, no differences were observed between any subgroup. For urgencies, the specificity for Black ethnicity (37.0%, 95% CI: 23.7–52.2) was significantly greater than the specificity for Asian (18.0%, 9.5–30.9) and for White (18.3%, 11.5–27.3) ethnicities and the specificity for IMD Q2 (32.0%, 26.7–37.7) was significantly greater than the specificity for Q4 (18.7%, 11.1–29.2) (Q2 is more deprived than Q4). No other significant difference was observed between any subgroup.

On EVal, the sensitivity for emergencies in the age group (70,+) was lower than the sensitivities for the other groups. There was an association between older age and late presentation, with 47.9% [36.2–59.6] of patients in this subgroup presenting after 1 week of symptoms onset, compared to 25.0% [18.1–33.3] in the subgroup aged [–,30]. Analysing subgroups with time of presentation <1 week and ≥1 week, there was no difference on sensitivities or specificities among different age groups. No differences existed in sensitivities or specificities for urgencies. All comparisons are available in the Supplementary material.

False negatives

All attendances in the TVal and EVal datasets which had a model-predicted priority that was lower than the labelled priority were considered potential serious misses.

On TVal, 45 (3.5%) potential serious misses occurred. Six emergency attendances (6/561 = 1.0%) were predicted as elective (Table 4). Twenty-five emergency attendances (25/561 = 4.5%) were predicted as urgent. Fourteen urgency attendances (14/179 = 7.8%) were predicted as elective.

On EVal, 20 (2.6%) potential serious misses occurred. Six emergency attendances (6/392 = 1.5%) were predicted as elective (Table 4). Thirteen emergency attendances (13/392 = 3.3%) were predicted as urgent. One urgency attendance (1/22 = 4.5%) was predicted as elective.

Discussion

We present the development, temporal (TVal) and external (EVal) validation of a ML-based system to optimise triage in ophthalmic emergency (DOTS).

	Sample sizes	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Proportion triaged in (%)	Confusion matrix	
Nurses (TVal)	Emergency = 561 (44.2) Urgency = 179 (14.1) Elective = 529 (41.7)	96.4 [94.5, 97.7]	25.1 [22.0, 28.5]	50.5 [47.5, 53.5]	89.9 [84.9, 93.5]	84.4 [82.3, 86.3]	541 530	20 178
Dots (TVal)	Emergency = 561 (44.2) Urgency = 179 (14.1) Elective = 529 (41.7)	94.5 [92.3, 96.1]	[a]42.4 [38.8, 46.1]	[a]56.5 [53.3, 59.7]	90.6 [87.0, 93.4]	[a]73.9 [71.4, 76.3]	530 408	31 300
Dots (EVal)	Emergency = 392 (51.5) Urgency = 22 (2.9) Elective = 347 (45.6)	95.2 [92.5, 97.0]	[a]32.2 [27.4, 37.0]	[a]59.9 [55.9, 63.7]	86.2 [79.4, 91.2]	81.9 [78.9, 84.5]	373 250	19 119

PPV: positive predictive value; NPV: negative predictive value; TVal: temporal validation; EVal: external validation. Data are % [95% CI]. [a]Significantly different from the proportion for NURSES (TVal) ($p < 0.05$).

Table 3: **DOTS performance compared to nurses for emergencies.**

DOTS is intended to help triage nurses to make more accurate and safe triage decisions, reducing inappropriate emergency attendances and helping streamline patients who can be safely managed in non-acute settings.

A large proportion of attendances to emergency departments were not actual emergencies, both at MEH (55.8%) and at UFMG (48.5%), and relatively simple attendances such as cases of dry eyes and blepharitis were among the most frequent ones. In MEH, triage nurses, even being highly skilled and trained, still decided that patients should be seen immediately in emergency even in most of elective attendances (3905/5437 = 71.8%).

On TVal, DOTS was as sensitive as nurses, but more specific (42.4% vs 25.1%), resulting in a higher PPV (56.5% vs 50.5%) and a lower rate of patients triaged to be seen immediately at emergency (73.9% vs 84.4%). For urgency, DOTS did not present any significant difference to the triage nurses' performance. These results indicate that DOTS has the potential to reduce the number of patients seen in the same day, reducing the emergency costs in a safe way.

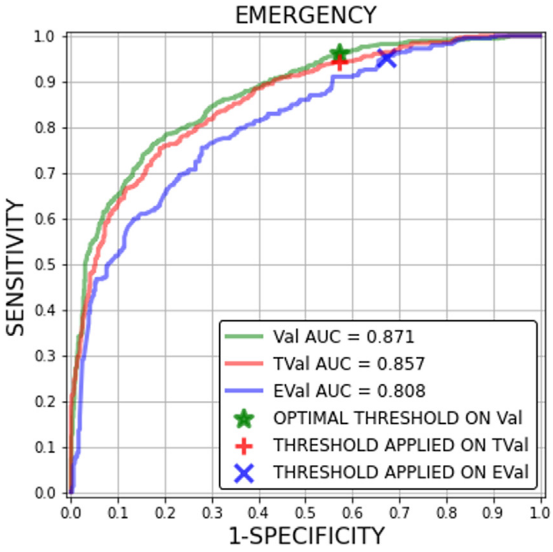

Fig. 2: ROC curves of DOTS with selected threshold for classifying emergencies in different datasets. Val: internal validation; TVal: temporal validation; EVal: external validation.

DOTS was externally validated using data inputs from triage nurses in a different setting and country. Compared to MEH dataset, the UFMG dataset presented a different distribution of diagnoses and presentations, with higher proportions of trauma and infective conditions and a lower proportion of complete patient histories. Even under these circumstances, DOTS was as sensitive as nurses from MEH, but more specific to emergencies (32.2% vs 25.1%). PPV, NPV and proportion of patients triaged in (positive rate) cannot be directly compared because they vary with disease incidences. However, the increase in specificity while keeping the sensitivity stable is expected to increase PPV and NPV, and to decrease the proportion triaged in.

The areas under the ROC curves were above 0.8 when applying DOTS to all the datasets analysed (Fig. 2). We notice that, for the region of interest of the application of the system (high sensitivity, right-hand side), the curves are similar in different datasets, almost overlapping up to a specificity of 0.3. When applying the threshold that was defined during internal validation (Val) to TVal and EVal, there was almost no shift in TVal operating point in relation to Val, and a small shift to the right in EVal operating point in relation to Val (similar sensitivity but lower specificity). These results highlight DOTS robustness in its desired operating region. It is important to note that even ROC curves with similar areas might have drastically different real-world performance if the thresholds shift under different settings.

To make transparent recommendations, all predictions are interpretable using SHAP values. All features observed are consistent with the current knowledge, such that eye injury/trauma, unilateral

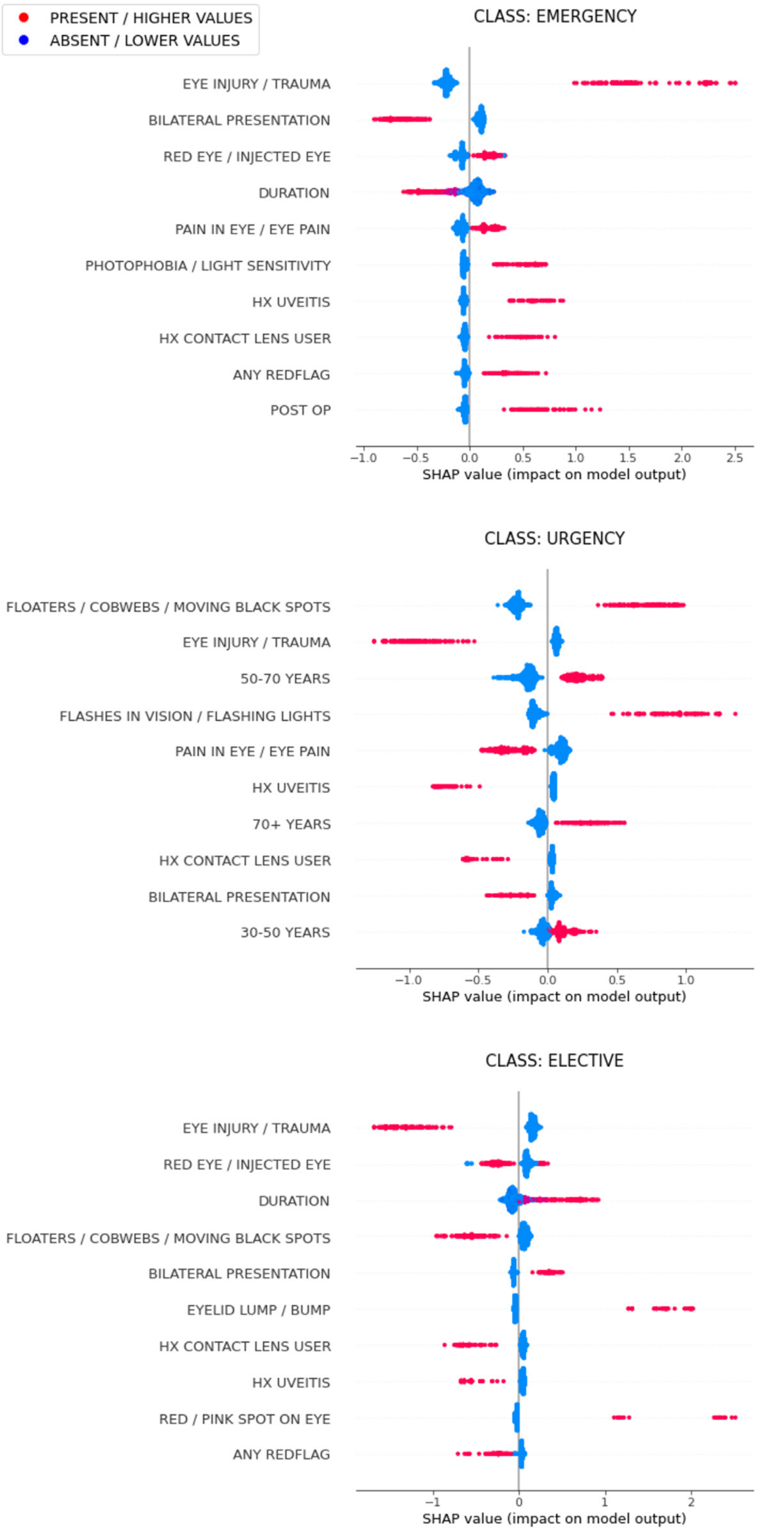

Fig. 3: Priority model explainability using SHAP values (validation). A feature with a positive SHAP value increases the likelihood that the model will make the relevant prediction. Features are ranked by order of importance from top to bottom. SHAP: shapley additive explanations; HX: history.

	TVal (6/561 emergencies = 1.0%)	EVal (6/392 emergencies = 1.5%)
Laterality	Bilateral: 5 (83.3) Unilateral: 1 (16.7)	Unilateral: 6 (100)
Duration	>4 weeks: 3 (50.0) 2–3 days: 2 (33.3) 1–4 weeks: 1 (16.7)	>4 weeks: 3 (50.0) 1–4 weeks: 1 (16.7) <24 h: 2 (33.3)
Red Flags	No Red Flags: 6 (100.0)	No Red Flags: 6 (100.0)
Signs and symptoms	Discharge from eye (transparent), Eye burning sensation: 1 (16.7) Eye itchiness/itchy/pruritus, Pain in eye/eye pain: 1 (16.7) Blurriness/blurred/blurry vision: 1 (16.7) Discharge from eye (yellow), Excessive lacrimation/tears/epiphora/watery eye, Red eye/injected eye: 1 (16.7) General symptom (headache): 1 (16.7) Discharge from eye (yellow): 1 (16.7)	Eye irritation/irritated: 1 (16.7) Eye itchiness/itchy/pruritus, Foreign body sensation/grittiness/feeling something in the eye: 1 (16.7) Pain in eye/eye pain: 1 (16.7) Pain in eye/eye pain, General symptom (headache): 1 (16.7) Foreign body sensation/grittiness/feeling something in the eye, Sticky eye/difficulty opening eyes on waking: 1 (16.7) Foreign body sensation/grittiness/feeling something in the eye, Excessive lacrimation/tears/epiphora/watery eye: 1 (16.7)
History	None/NA: 3 (50.0) General ophthalmology service patient: 1 (16.7) Ocular comorbidity (other): 1 (16.7) Glaucoma and Uveitis services patient: 1 (16.7)	None/NA: 6 (100.0)
Diagnoses	Anterior uveitis, staphylococcal hypersensitivity: 1 (16.7) Blepharitis, lacrimal apparatus inflammation/infection: 1 (16.7) Papilloedema: 1 (16.7) episcleritis or scleritis, staphylococcal Hypersensitivity: 1 (16.7) Other retinopathy: 1 (16.7) Lacrimal apparatus inflammation/infection: 1 (16.7)	Conjunctival/corneal injury: 4 (66.7) Extraocular herpes: 1 (16.7) Blind painful eye: 1 (16.7)
Nurse triage outcome	See in emergency: 4 (66.7) See in UCC: 2 (33.3)	Green (4th priority): 5 (83.3) Blue (5th priority): 1 (16.7)

TVal: temporal validation; EVal: external validation; NA: not available.

Table 4: **Summary of false negatives on TVal and EVal: emergency predicted as elective.**

presentation, lower duration, red/injected eye, and eye pain are the five most important inputs for increasing the chance of a case to be predicted as an emergency. The confirmation of a sensible interpretation increases our confidence on the quality of the data and of the model.

There are increasing disparities in the representation of population and disease groups in ophthalmic image databases.[17] Furthermore, models trained on datasets of skewed populations could engrain systematic biases and reinforce previous undesirable practice.

Our tabular dataset used to build DOTS consists of presentations to Europe's largest eye emergency department over a 1-year period encompassing anterior and posterior segment pathologies.[3] We conducted bias analyses in demographic subgroups within TVal and EVal. Our demographic data reflects the difficulty of collecting patient ethnicity in healthcare settings, with more than 58% of attendances in TVal being registered in the system as of "other" or "unknown" ethnicities. Therefore, performing subgroup analysis for fairness based only on ethnicity data may be misleading not only due to the possible poor correlation between ethnicity and socio-economic indicators but also due to possible low data quality. On the other hand, less than 1% of IMD values were missing, and our fairness analysis using IMD is more robust.

On TVal, no differences were observed in sensitivities for emergency or urgency in subgroups, indicating that the model was equally safe in them. Similarly, no differences were observed in specificity for emergency, therefore false positive referrals were not biased by patient demographics. For urgency, the higher specificities observed for Black ethnicity and for IMD Q2 in relation to Q4 (Q2 is more deprived than Q4) indicate that DOTS performed better (less false positives without more false negatives) in those subgroups that can be considered as groups at risk. Considering that black people experience delays in accessing care in emergencies, the model performance observed has important implications for closing, or at least not increasing, the clinical outcome gap between patients of White and minority ethnicities and low vs high socioeconomic status.[18]

On Eval, the sensitivity for emergencies in the age group (70,+) was lower than the sensitivities for the other groups. However, there was a larger proportion of patients with late presentations in this group, with almost half of them presenting after 1 week of symptoms onset. Differences in sensitivities disappeared when subgroups with time of presentation <1 week and ≥1 week were analysed separately, confirming our hypothesis that presentation time was a confounding factor in the former analysis. No differences existed in sensitivities or specificities for urgencies. Even with the possible explanation of differences in sensitivity being due to presentation time, what is probably a bias of the healthcare system and not of DOTS, we should keep monitoring for biases when DOTS is deployed.

DOTS was found to be robust to incidence changes when tested on the simulated scenarios. Even with large variations in disease incidences, for emergencies, sensitivities were always greater than 92% and specificities were invariably greater than 36%. No significant differences in performance were observed between any of the scenarios and the original TVal, either for emergency or urgency. These findings suggest that DOTS is relatively robust to "dataset shift", a major cause of failure in ML, where a model underperforms because the dataset on which it is trained becomes outdated.[19]

We looked for and disclosed potential serious misses both on TVal (3.5%) and on EVal (2.6%). It is important to note that not all attendances flagged as potentially serious misses are actual misses, since their definition was based on diagnoses reported on the EMR that can be, for example, previous diagnoses.

Among the emergency cases predicted as elective, on TVal, 67% presented symptoms >1 week, with no red flags and only including non-specific symptoms such as eye discharge, burning, itchiness and blurry vision. On EVal, two thirds of the cases were conjunctival/corneal injuries with no history of trauma registered. Therefore, with the given inputs, even experienced ophthalmic consultants would agree that several of the cases are not emergencies.

This study has some limitations. Firstly, most of its data comes from a single tertiary ophthalmic centre (MEH). However, when DOTS was externally validated in a different continent and in simulated scenarios with varied disease incidences, it had consistently shown non-inferior performance. This suggests that DOTS is robust to incidence changes, assuming clinical features of specific diseases do not vary across sites.

Secondly, the study did not assess human-computer interaction (HCI) which potentially impacts triage accuracy. For instance, the user may be inclined to triage patients to the higher priority between their clinical suspicion and model prediction, thus leading to overall increased emergency referrals. Contrarily, the user may triage to the lower priority between clinical suspicion and model prediction, thus increasing the risk of serious misses. Moreover, holistic considerations of care, such as safeguarding concerns in vulnerable patients and quality of life in patients with only one eye, may impact the decision for a patient to be seen urgently. Future HCI analysis is needed to assess impact on triage outcome and corresponding safety net measures.

Thirdly, automation bias may develop over time whereby the user relies on the prediction generated. Over reliance on the system, clinician de-skilling, and workflow disruption by technological failure are intrinsic risks; worse still, these may overspill to clinical errors and patient harm.[20] Continuous mechanisms for system safety control and future research are needed to assess those risks in real-world deployment.

Finally, the nurses who participated in the study were not characterised by their triage experience which potentially was a confounding for triage performance. Whilst this was mitigated by the 1-year data collection period at MEH, future studies could compare model performance with triage nurses of varied triage experience.

DOTS will be tested in a single-site clinical trial at MEH emergency. The clinical trial will consider user triage experience, provide insights on HCI and real-world performance, and test and improve safety control mechanisms.

In this study, we developed and validated DOTS, an ML-based system to support triage nurses in ophthalmic emergency, using data from over 13,000 emergency attendances. In temporal validation, DOTS presented equivalent sensitivity to triage nurses to determine attendance priority, but a specificity 17.3% higher, so identifying 10.5% fewer patients, who did not need to be seen immediately at emergency. In external validation, DOTS presented equivalent sensitivity to triage nurses at MEH to determine attendance priority and was 7.1% more specific to emergencies. DOTS showed consistent performance levels in social-demographic subgroups, being fair, and robust to varying disease incidences in simulated scenarios. A clinical trial will be needed to validate the findings of this study and provide insights on real-world deployment. Such a clinical trial is planned.

Contributors

CBR, MM, and ACD were involved in the original study conception. MM, ACD and GRH were involved in the funding acquisition. MM, ACD were involved in the project administration. CBR, MM, EL, AJ, YNN, PC, VR, GV, and MLH were involved in data acquisition and curation. CBR, MM, AJ, ACD, and EL accessed and verified the underlying data from MEH. CBR, JRF, PC, VR, GV, MLH, and DVS accessed and verified the underlying data from UFMG. CBR developed the models. CBR and JRF analysed the data. CBR and EL wrote the initial draft of the manuscript. ACD, DVS, PAK, AJ, YNN, MM and PS reviewed and edited the manuscript. MM, ACD, AJ, DVS, PAK, and GRH supervised the study. All authors reviewed and approved the final draft of the manuscript.

Data sharing statement

Deidentified data generated and analysed during the study were acquired through data use agreements with MEH. Requests for data will require approval from MEH and the signing of data access agreements; requests can be submitted to the corresponding author.

Declaration of interests

CBR, MM, EL, and JRF work for DemDX Ltd. CBR, MM, AJ, EL, YNN, and ACD were funded by the National Institute for Health and Care Research (NIHR) artificial intelligence (AI) Award to work on this study. PAK has acted as a consultant for Roche, Novartis, Boehringer-Ingleheim, Adecco, Bitfount and is an equity owner in Big Picture Medical. He has received speaker fees from Novartis, Gyroscope, Bayer, Thea, Boehringer-Ingleheim, Apellis, Abbvie, Alimera, Roche, Genentech, Specsavers, Heidelberg Engineering, Topcon, and Santen. He has received travel support from Bayer, Topcon, and Roche. He has attended advisory boards for Boehringer-Ingleheim, RetinAI, Novartis, Apellis, Abbvie, and Roche. GRH, PS, PVIPC, VOMR, GFSV, MLOH, and DVVS declare no competing interests.

Acknowledgements

This research was supported by the NIHR Moorfields Clinical Research Facility and NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology. It was funded by the Artificial Intelligence in Health and Care Award (AI_AWARD01671) of the NHS AI Lab under National Institute for Health and Care Research (NIHR) and the Accelerated Access Collaborative (AAC). PAK is supported by a Moorfields Eye Charity Career Development Award (R190028A) and a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1). We thank all triage nurses at the Accident and Emergency Department at Moorfields Eye Hospital NHS Foundation Trust, London, UK and all staff at the Accident and Emergency of Hospital Sao Geraldo, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil. We thank the Moorfields IT team and patients on the study Lay Advisory Panel for their feedback on the study. The views expressed in this publication are those of the authors and do not necessarily those of the NIHR or the Department of Health and Social Care.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.102331.

References

1 Resnikoff S, Lansingh VC, Washburn L, et al. Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? *Br J Ophthalmol*. 2020;104(4):588–592. https://doi.org/10.1136/bjophthalmol-2019-314336.

2 NHS Digital. Hospital outpatient activity 2018-19. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity/2018-19. Accessed October 3, 2023.

3 Smith HB, Daniel CS, Verma S. Eye casualty services in London. *Eye Lond*. 2013;27:320–328.

4 NHS Digital. Hospital accident & emergency activity 2020-21. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident–emergency-activity/2020-21. Accessed October 2, 2023.

5 Channa R, Zafar SN, Canner JK, Haring RS, Schneider EB, Friedman DS. Epidemiology of eye-related emergency department visits. *JAMA Ophthalmol*. 2016;134(3):312–319. https://doi.org/10.1001/jamaophthalmol.2015.5778.

6 Baker LC, Baker LS. Excess cost of emergency department visits for nonurgent care. *Health Aff Millwood*. 1994;13(5):162–171. https://doi.org/10.1377/hlthaff.13.5.162. Winter.

7 Li YYS, Vardhanabhuti V, Tsougenis E, Lam WC, Shih KC. A proposed framework for machine learning-aided triage in public specialty ophthalmology clinics in Hong Kong. *Ophthalmol Ther*. 2021;10(4):703–713. https://doi.org/10.1007/s40123-021-00405-7.

8 Chen J, Wu X, Li M, et al. EE-explorer: a multimodal artificial intelligence system for eye emergency triage and primary diagnosis. *Am J Ophthalmol*. 2023;252:253–264. https://doi.org/10.1016/j.ajo.2023.04.007.

9 Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1–W73. https://doi.org/10.7326/M14-0698.

10 BMJ best practice. https://bestpractice.bmj.com/info/. Accessed October 3, 2023.

11 NHS Digital. Ethnicity - part of data quality of protected characteristics and other vulnerable groups. https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set/submit-data/data-quality-of-protected-characteristics-and-other-vulnerable-groups/ethnicity. Accessed October 3, 2023.

12 Ministry of Housing, Communities & Local Government. National statistics - english indices of deprivation. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019; 2019. Accessed October 3, 2023.

13 Amazon Web Services, Inc. https://aws.amazon.com. Accessed October 3, 2023.

14 Van Rossum G, Drake FL. *Python 3 reference manual*. CreateSpace; 2009.

15 Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *Inf Fusion*. 2022;81:84–90. https://doi.org/10.1016/j.inffus.2021.11.011.

16 Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.

17 Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health*. 2021;3(1):e51–e66. https://doi.org/10.1016/S2589-7500(20)30240-5.

18 Wilper AP, Woolhandler S, Lasser KE, et al. Waits to see an emergency department physician: U.S. trends and predictors, 1997-2004. *Health Aff*. 2008;27(2):w84–w95. https://doi.org/10.1377/hlthaff.27.2.w84.

19 Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283–286. https://doi.org/10.1056/NEJMc2104626.

20 Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. *J Med Internet Res*. 2023;25:e43251. https://doi.org/10.2196/43251.