tvst

Special Issue

Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening

Yuchen Xie[1,*], Dinesh V. Gunasekeran[1,2,*], Konstantinos Balaskas[3], Pearse A. Keane[3], Dawn A. Sim[3], Lucas M. Bachmann[4], Carl Macrae[5], and Daniel S. W. Ting[1,6,7]

[1] Singapore National Eye Center, Singapore Eye Research Institute, Singapore
[2] School of Medicine, National University of Singapore, Singapore
[3] Moorfields Eye Hospital, National Health Service, London, UK
[4] Clinical Epidemiology, University of Zurich, Zurich, Switzerland
[5] Business School, Nottingham University, Nottingham, UK
[6] School of Medicine, Duke-National University of Singapore, Singapore
[7] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, Guangdong, China

Correspondence: Daniel S.W. Ting, Duke-National University of Singapore Medical School, 11 Third Hospital Avenue, Singapore 168751. e-mail: daniel.ting.s.w@singhealth.com.sg

Received: January 5, 2020
Accepted: January 23, 2020
Published: April 13, 2020

Keywords: artificial intelligence; ocular imaging; diabetic retinopathy; machine learning; deep learning

Citation: Xie Y, Gunasekeran DV, Balaskas K, Keane PA, Sim DA, Bachmann LM, Macrae C, Ting DSW. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. Trans Vis Sci Tech. 2020;9(2):22, https://doi.org/10.1167/tvst.9.2.22

Systematic screening for diabetic retinopathy (DR) has been widely recommended for early detection in patients with diabetes to address preventable vision loss. However, substantial manpower and financial resources are required to deploy opportunistic screening and transition to systematic DR screening programs. The advent of artificial intelligence (AI) technologies may improve access and reduce the financial burden for DR screening while maintaining comparable or enhanced clinical effectiveness. To deploy an AI-based DR screening program in a real-world setting, it is imperative that health economic assessment (HEA) and patient safety analyses are conducted to guide appropriate allocation of resources and design safe, reliable systems. Few studies published to date include these considerations when integrating AI-based solutions into DR screening programs. In this article, we provide an overview of the current state-of-the-art of AI technology (focusing on deep learning systems), followed by an appraisal of existing literature on the applications of AI in ophthalmology. We also discuss practical considerations that drive the development of a successful DR screening program, such as the implications of false-positive or false-negative results and image gradeability. Finally, we examine different plausible methods for HEA and safety analyses that can be used to assess concerns regarding AI-based screening.

Introduction

Ageing populations and other demographic shifts have made diabetes mellitus (DM) a major epidemic of the 21st century.[1] Despite improvements in health care that have led to decreasing age-specific mortality worldwide,[2] there is an increasing net burden of DM disease, lives lost, and years lived with the disease and its complications.[1] At a prevalence of over one-third of patients with DM,[3] diabetic retinopathy (DR) is one such disabling complication of the disease that now presents a mounting challenge to over-stretched eye care services worldwide.[3–5] The need for regular screening has been established for early detection of DR along with diabetic macular edema (DME)[6]; whereby DME can develop anytime as DR progresses,[7] and is a frequent cause of severe visual impairment in these patients.[8]

These trends give rise to an urgent need for solutions that can sift through the growing crowds of at-risk individuals and triage patients in need of early treatment to prevent permanent vision loss.[9,10] Fortunately, progress in the parallel fields of ophthalmic imaging and the deep learning (DL) branch of artificial intelligence (AI) have enabled promising solutions that automate the detection of major blinding eye diseases in ophthalmic images.[10] Automation of screening using AI-based solutions could thereby free up limited health care resources to provide more complex eye care services, cater to subpopulations with barriers to health care access, or facilitate transition from opportunistic to systematic screening programs.

Clinically acceptable performance of these AI-based solutions in health care has been established for the application of DR screening based on classification of ophthalmic imaging, with area under the receiver operating characteristic curve, sensitivity, and specificity in excess of 80%.[9] These solutions thereby enable accurate diagnosis of disease severity for triage and right siting of patients.[10,11] However, despite the increasing interest in these AI-based solutions, few have been implemented across populations owing to uncertainty regarding their application to different health care settings, as well as the potential safety challenges.[10,12,13] In this article, we highlight the major challenges in integrating AI-based solutions for DR screening, along with considerations for conducting health economic assessment (HEA) and safety analysis of these solutions.

Applications of AI in DR Screening

The clinical features of DR in the retina that are indicative of clinical severity and outcomes (e.g., blindness) have been described in the existing literature and consolidated in clinical guidelines, such as the International Clinical Classification of Diabetic Retinopathy Scale.[14] This body of knowledge has fueled applications of AI in the form of classical feature-based image analysis and machine learning (ML) algorithms for DR screening training based on individual feature labeling by experts. These methods have been successfully used to automate classification of retinal fundus photographs based on the presence/absence (binary classification) and/or clinical severity (multiclassification) of DR.[15]

The advent of deep learning systems (DLS) heralds a new era in the processing of medical data using AI, whereby algorithms are trained on large repositories of imaging data without individual feature labeling. Instead, training is conducted using imaging data with labeling of overall clinical severity by experts. The DLS then self-learns predictive features from these labels using mathematical functions.[10] Recent reports of DLS outperforming the classical feature-based image analysis in screening for DR and other ocular diseases have been described.[10,16]

The development and validation of several novel DLS solutions for automated DR screening have been reported by groups from various countries, including Singapore, United States, United Kingdom, China, Thailand, India, and Africa.[17–23] These investigators reported clinically acceptable performance of their DLS tools for classifying DR in color fundus photography or optical coherence tomography (OCT) imaging.[9] Some AI-based solutions for DR screening have been approved as medical devices for automated classification of ophthalmic imaging based on evidence from studies conducted in several high-income countries. These solutions also have tremendous potential to enhance health care in resource-limited settings.[18,24]

Methods of HEA

Given the scarcity of resources available within a health system, HEA of novel health technologies is required for decision-makers to efficiently allocate resources. DR screening and teleophthalmology programs are cost-effective in a variety of developed[25–31] and developing settings.[32,33] Before the advent of DL, feature-based computing techniques had been developed for automated retinal image screening.[20,34–36] Such automated retinal screening has been shown to be cost-effective when applied to the national screening program in Scotland and the United Kingdom.[34–37]

However, few studies have incorporated HEA for teleophthalmology services augmented with DL-based classifiers for DR screening. The few existing reports on HEA based on the implementation of DL-based solutions for DR screening are from the United Kingdom and Singapore. These studies show AI to be cost-effective in Singapore and the United Kingdom. However, this finding may not be generalizable given that they are both high-income countries with established teleophthalmology DR screening program. Cost-effectiveness may differ between countries owing to variations in disease prevalence, geographic barriers, availability/cost of the relevant skilled manpower, and health care resources. There have not been any studies conducting HEA of AI applications for DR screening in resource-limited settings, or countries without established teleophthalmology DR screening programs to date.

The most appropriate HEA method for a given test or intervention is determined by several factors depending on the existing evidence for the solution and the intended clinical context for its application. Given the relatively nascent nature of AI-based solutions for health care, there is a need to identify suitable methods of HEA to evaluate them. In the following section, we outline common types of HEA and the contexts in which they are applied. These include

Table 1. Types of HEA

Method	Measurement of Effect	Questions Raised	Measurement of Cost
CUA	Healthy years (typically measured as quality-adjusted life years)	Given financial constraints, what is the most efficient way of allocating limited resources for improved outcomes?	Monetary units
CEA	Natural units (e.g., life years gained, cases of blindness avoided, and others)	Given financial constraints, what is the most efficient way of allocating limited resources for improved outcomes?	Monetary units
CMA	Assumption is that the clinical effectiveness of each alternative is the same	Given a certain objective, what is the most efficient way to achieve it?	Monetary units
CBA	Monetary units	Should a given goal or objective be pursued and to what extent?	Monetary units

cost-effectiveness analysis (CEA), cost-utility analysis (CUA), cost-minimization analysis (CMA), and cost-benefit analysis (CBA) (Table 1).[38,39]

Cost-Utility/Cost-Effectiveness Analysis

CEA and CUA are two distinct forms of HEA that are often used interchangeably in the literature,[38] although CUA is technically more comprehensive. CEA generally uses a single clinical outcome (life years), whereas CUA often uses quality-adjusted life years (i.e., calculated based on preferences for a particular health state).[38–40] Health Economics authorities (e.g., Washington Panel[41] and the official requirements of economic evaluations of the United Kingdom[42]) have recommended the use of CUA. When conducting CEA, other clinical outcomes (e.g., case of blindness avoided or cases of DR detected) can be used instead based on the disease studied.[38] Examples of other outcomes that may be relevant to DR screening are blindness cases averted in a primary health care setting,[43] or number of cases of proliferative DR detected in a screening network.[28]

Cost-Minimization Analysis

CMA is often used when it has been established that two or more health technologies/interventions have comparable clinical effectiveness. In this context, researchers are primarily interested in assessing which alternative is less costly and quantifying the potential saving associated with the least expensive alternative.[44] However, one of the major concerns with CMA is that it is often difficult to establish whether two alternatives are indeed equivalent (e.g., in a longitudinal study).[38] Several researchers argued that even when there is no statistical difference found between the effectiveness of the two alternatives (i.e., no statistically significant difference in clinical outcomes), CEA is still preferred for HEA.[45–47]

CMA is primarily used in situations with an established expert consensus (e.g., professional, based on research) that the two alternatives are equivalent in clinical effectiveness.[48] It has been suggested that CMA is most suitable for clear-cut scenarios when alternatives represent similar state-of-the-art solutions (e.g., screening tools of the same class).[45,48] A research practice report by the International Society for Pharmacoeconomics and Outcomes Research indicated that CMA provides useful insights on budget impact for decision-makers.[46]

In DR screening, the current literature has concluded that AI-based solutions using DL techniques have demonstrated clinically comparable performance to human assessment in established DR screening programs, both in publicly available datasets and real-world settings.[10] A recent meta-analysis on DLS published has further confirmed this conclusion.[11] As such, CMA is a viable method to conduct HEA of comparable AI-based solutions in health systems with established DR screening programs.

Cost-Benefit Analysis

CBA is often used to quantitatively evaluate whether a new intervention should be adopted by directly comparing costs of the intervention against existing practices. For CBA, clinical outcomes and effects (e.g., disability days avoided, life years gained, medical complications avoided, or quality-adjusted life years gained) are converted into monetary value to evaluate the foreseeable net costs of adopting a given solution

Table 2. Health Economic Studies on DR Screening Using AI

Author, Year, Country	Comparators	Screening Model	Measurement of Effect	Economic Outcomes
Scotland et al,[34] 2007, UK	Semi-automated grading (hybrid approach) vs. manual grading alone	Digital photography and multilevel manual grading systems	The number of appropriate screening outcomes (i.e., defined as final decisions appropriate to actual grade of retinopathy present) and true referable cases detected in one year	Compared to the manual grading model, the semi-automated model led to a saving of £4088 per additional referable case detected, and of £1990 per additional appropriate screening outcome.
Tufail et al,[20] 2016, UK	AI-based ML tool as placement for initial manual grading (semi-automated hybrid)	AI-based (ML) two-field fundus photos	Appropriate outcomes (defined as identification of DR present vs. absent by the AI-based software)	AI-based semi-automated hybrid approach (Retmarker and EyeArt) had sufficient specificity to make them cost-effective to manual grading alone, as ICER was $18.69 and $7.14, respectively
Xie et al,[50] 2019, Singapore	Semi-automated hybrid approach (DLS-based) vs. manual grading alone	Retinal fundus photographs	QALYs	DLS-based (semi-automated hybrid approach) resulted in a lifetime cost-saving of $135 per patient while maintaining comparable QALYs gained.

QALYs, quality-adjusted life years;
ICER, incremental cost-effectiveness ratio;
manual grading is equivalent to human assessment.

within a clinical pathway.[48] The difficulty of applying CBA effectively in health, however, is the difficulty in assigning a monetary value to clinical outcomes (e.g., quality-adjusted life years or blindness prevented).[39]

In discrete choice experiments, patients are invited to express their strength of preference based on specific clinical outcomes to help ascribe a monetary value to them. However, this is subject to variation from cultural differences, and there are challenges (e.g., uncertainty about the validity of the outcomes of interventions) that need to be addressed for CBA to be used in HEA of AI-based solutions for DR screening. Cartwright[49] has contributed an insightful review of several reports applying CBA to the intervention of drug abuse treatment services. Notably, they highlighted challenges in the measurement of clinical outcomes, need for representative populations of patients recruited, and lack of standardization in the application of CBA.

HEA of AI-Based Solutions for DR Screening

The previous section indicated that the existing reports of HEA of AI-based solutions for DR screening are from countries with established teleophthalmology programs, and systems for training and regular examination of human assessors for DR screening (i.e., United Kingdom and Singapore). Having reviewed these reports, one would arrive at the conclusion that semi-automated screening models are cost-effective (Table 2).[20,34,50] Tufail et al.[20] reported the cost saving to be 12% to 21% for DR screening in the United Kingdom using ML (an AI-based technology) in comparison with human assessors.[17,20] A Scottish study showed a 46.7% cost-reduction by replacing first-level human assessment with automated grading in a national DR screening program.[34,37] A study from Singapore suggested that the semi-automated/AI-assisted screening model is cost-effective compared to human assessment for DR screening over a lifetime horizon.[50] However, there is no published HEA of a fully-automated DR screening model to date.

Implications of False Negatives (FNs) in Screening Programs

FN cases are patients with referable DR that are mislabeled as being normal. As a result, these patients may receive delayed care if they are only referred at a subsequent screening interval that could be months or years later. The clinical impact of delayed care is the risk of interim disease progression.[51] For DR this can lead to permanent vision loss in severe cases, as they tend to progress faster.[52] Even when effective treatment is readily available, a high FN rate puts patients at increased risk of disease progression and vision loss.[53,54] Notwithstanding the financial burden on the health care system from disease progression due to late detection, studies also report a psychological impact on patients, loss of public confidence in screening programs, and legal implications as other major consequences of FNs.[55]

Implications of False Positives (FPs) in Screening Programs

In contrast to FN, a high FP rate of screening programs results in referrals of normal screening subjects for further assessment by an ophthalmologist when it is not required. This will create additional costs for the health care system in terms of resources and manpower being utilized to attend to unnecessary referrals. Moreover, FPs from a screening program could result in unnecessary anxiety and psychological stress for patients.[54] However, there is no expert consensus on the acceptable FP rate performance for DR screening to date.[56]

Image quality is another important consideration in real-world screening implementation. Images with low quality (ungradable images) would be referred to the assessors and could incur additional costs for regrading images or repeat image acquisition if necessary. Nevertheless, the treatment of ungradable images as FP is not yet standard practice. In reports of DL models, several groups have excluded ungradable images from their analyses.[18,54] However, this may not reflect the true performance of these solutions in practical application.

In a study of automated eye screening, Tufail et al.[20] reported results after including images of poor quality or classified as ungradable by the human assessors. Similarly, Ting et al.[16] in Singapore also considered ungradable as referable DR to avoid missing possible DR cases. Discrepancies in reporting FP rate would impact the HEA of screening programs. The authors recommend that images classified as ungradable by AI-based solutions for DR screening should be included in the assessment of performance to reflect the practical need for these patients to be referred for definitive assessment.

In developing a screening solution, there is a trade-off between minimizing for FP or FN. The ideal balance for each health system may vary slightly depending on their system factors, such as cost structures, availability of resources, as well as resolution of competing clinical and financial interests. However, when clinical considerations are prioritized, minimizing FN in the context of these high performing AI-based solutions is generally favored because of the potential clinical safety impact of FN, whereas that for FP is mitigated when patients are reviewed by the attending ophthalmologist.

Challenges of Conducting HEA in the Real-World Setting

A number of recent studies suggest that the use of traditional techniques for HEA to quantify the impact of complex health services, such as a national screening program, can be challenging.[57–61] They explained that the evaluation of complex interventions involving both human services and advanced assistive technologies will likely encounter a number of problems. Among them, the heterogeneity of the user groups, participant selection (bias), the degree of participation of the user groups carrying out the intervention, and the composition of these groups lead to complexities that may require modifications to traditional assessment methods.[59]

In addition, conducting a comprehensive evaluation of an AI-based solution for DR screening requires consideration of local context, such as the availability of skilled manpower and DR screening resources. Therefore investigating the implementation in resource-limited settings is also an important area for future health services research. This is needed to evaluate the interventions in these settings based on their unique practical considerations, such as limited availability of internet access and the forms of imaging devices available (table-mounted, handheld, smartphone adapter-based, and others) that may affect image quality and the performance of AI-based solutions for DR screening, such as the incidence of FPs and FNs.[24]

Summary Recommendations for HEA of AI-Based DR Screening

In summary, the choice of the specific HEA method for a particular clinical application of AI would depend on the form of application, clinical outcomes relevant to the intervention, availability of preexisting representative datasets, and the nature of assumptions associated with the solution. CMA is useful for rapid comparison of interventions with established comparable clinical effectiveness. Where this has yet to be established, CUA is often the preferred mode of analysis, although the nature of measurable clinical outcomes may require CEA to be considered instead. In resource-limited settings with high unmet clinical needs, CBA provides a tool for quantitative assessment of interventions to identify the most financially prudent option.

Based on these considerations, CMA can be considered to evaluate AI-based solutions for DR screening in developed countries with established DR screening programs. CEA/CUA may need to be conducted for other dissimilar contexts to evaluate both clinical outcomes and costs based on the health system in question. Given the pressing need for solutions to expand the capacity of DR screening capabilities, HEA using data from clinical trials would be ideal to provide reliable and timely results with high internal validity to aid administrators in decision-making regarding the adoption of these AI-based solutions.[60] The selected HEA method needs to be applied with established HEA strategies such as use of multiple comparator groups, stratified sensitivity analysis using those groups, and appropriate modeling methods, as outlined in frameworks for the assessment of complex public health interventions.[61]

Methods of Safety Analysis for Health Care

Implementing AI technologies in national screening programs have the potential to improve patient safety by providing rapid and reliable identification of referable eye disease. It also has the potential to introduce new risks that will need careful analysis and management.[12,62] These risks can be associated with the underlying AI technologies or the organizational systems that implement them. For example, mismatches can develop between the data that a DLS was originally trained on (i.e., training dataset) and the data it is required to interpret (validation dataset), such as geographic variations in disease phenotypes, which can lead to shifts in screening performance.[63]

Therefore organizational systems and decision-making processes need to be developed for periodic monitoring to investigate and address instances in which the automated screening system does not provide an appropriate classification to ensure that the overall screening system can "fail safe." In addition, analyzing the safety of a DLS can be challenging, owing to difficulties in understanding the underlying decision-making process. The safety analysis of AI-based screening programs therefore requires the use of analytic techniques that consider clinical, technical, social, and organizational sources of safety and risk. A range of safety analysis methods have been developed for the prospective analysis of potential risks in complex sociotechnical systems. However, there has been limited examination of how these can be applied to large-scale AI systems in health care to date. In the following sections, we outline several relevant methods of safety analysis, including failure mode and effects analysis (FEMA), system-theoretic process analysis (STPA), and bowtie tie analysis.

Failure Mode and Effects Analysis

FMEA is a structured and proactive approach to identifying safety issues in complex sociotechnical systems that is increasingly applied to health care.[64–66] FMEA involves creating a detailed map of processes for a service or activity to identify all the potential manners that those processes might fail, and what the causes and effects of those failures might be. Each failure is then assessed according to the severity of the outcome, the probability of occurrence, and the likelihood of detection, to prioritize mitigating action and resources.

One of the key requirements of conducting an effective FMEA is to establish a team with deep and broad expertise in all aspects of the system being analyzed, encompassing clinical, technical, and organizational components.[67] Conducting FMEAs can be time-consuming and resource intensive. Because of the focus on analyzing individual failure modes, capturing complex interactions between different parts of a system is also a challenge. However, FMEA provides a systematic approach to understand and develop solutions for a broad range of technical and organizational safety risks and could be effectively applied to the implementation of AI-based screening programs.[68]

System-Theoretic Process Analysis

STPA is a safety analysis method that analyses the way safety is controlled within a complex system, such as through automated monitoring, management supervision, or regular audits. It identifies where potential gaps in those control systems may occur, and how serious those unsafe control actions might be.[68,69] One of the core premises of this approach is that all systems have hierarchical control structures: for example, local-level control might be performed by technicians or clinicians; higher-level supervision may be conducted by program managers; and overall oversight may be performed by systems regulators.

The STPA method seeks to identify hazards in terms of potential failures of control, such as scenarios in which clinicians may not become aware of ungradable images. STPA is a relatively new method that requires extensive expertise in systems-analysis. It has seen limited application in health care to date, although its associated incident analysis model has been applied with useful outputs.[70,71] STPA may be particularly valuable in identifying and optimizing the safety monitoring and governance systems required for AI-based screening programs. These may include routine algorithmic audits, peer review, and adjudication processes, which have already been described as solutions for grader variability when training automated solutions for DR screening.[72]

Bowtie Analysis

Bowtie analysis is a barrier-based approach to safety analysis that is widely used in highly automated safety-critical industries, such as aviation, and is beginning to be applied in health care.[71,73,74] It provides a visual method to identify and map factors that contribute to a particular failure, the consequences that can result from that failure, and the barriers and risk controls that can protect against those contributing factors and consequences. One of the main strengths of bowtie analysis is the ability to produce comprehensive graphic representations of complex models of risk, which can be used to explore both the sources of risk and safety in relation to specific types of failure. Directly identifying safety barriers and risk controls also provides practical insight into the actions that are needed to mitigate risks when implementing a new system.[75]

Conducting Safety Analyses of AI-Based Solutions for DR Screening

In the earlier section, we have reviewed several important methods (FMEA, STPA, bowtie analysis) that can be used to analyze the safety concerns in implementing AI-based solutions in ophthalmology. To use these methods for safety analyses, a thorough understanding is required of the various potential models that AI-based solutions for DR screening can be implemented within a health system. The use of AI with teleophthalmology has been suggested as a sustainable solution to rapidly scale-up DR screening.[10,13]

Existing teleophthalmology screening programs utilize remote human assessment (by manual graders) to identify the presence of DR in ophthalmic imaging captured in community-based settings. To deploy AI-based DR screening programs, there are two different models that could be used: the semi-automated (using DLS as a filter prior to human assessment), and the fully-automated (using DLS as a complete replacement of human assessment).[16] Figure 1 depicts the two DLS-based DR screening models (Figs. 1B, 1C), alongside an existing teleophthalmology human assessment model (Fig. 1A).

The semi-automated model (Fig. 1B) is a hybrid approach using an AI-based solution as a preliminary filter prior to human assessment. Here referable cases from the solution undergo secondary assessment by human assessors in a centralized reading center. Cheung et al.[76] suggested that the benefits of the semi-automated model include decreased workload on nonreferable retinal images, and reduced FP cases referred to ophthalmologists. However, a fully-automated model (Fig. 1C) with complete replacement of human assessment may be more relevant for countries without existing systems and manpower for teleophthalmology. Ultimately, the manner in which various AI-based solutions are to be integrated into different health care systems needs to be considered based on the performance of the tool, the constraints of the system, and the safety considerations for participating patients.

Because of the scalability of AI and ability to meet the needs of varied populations of patients, there is growing interest to examine potential safety issues that need to be considered.[12] The earlier-mentioned methods for safety analyses can be used to inform the development of regulatory standards for assessment of safety and efficacy that are still evolving with the advances of AI applications in medicine.[77]

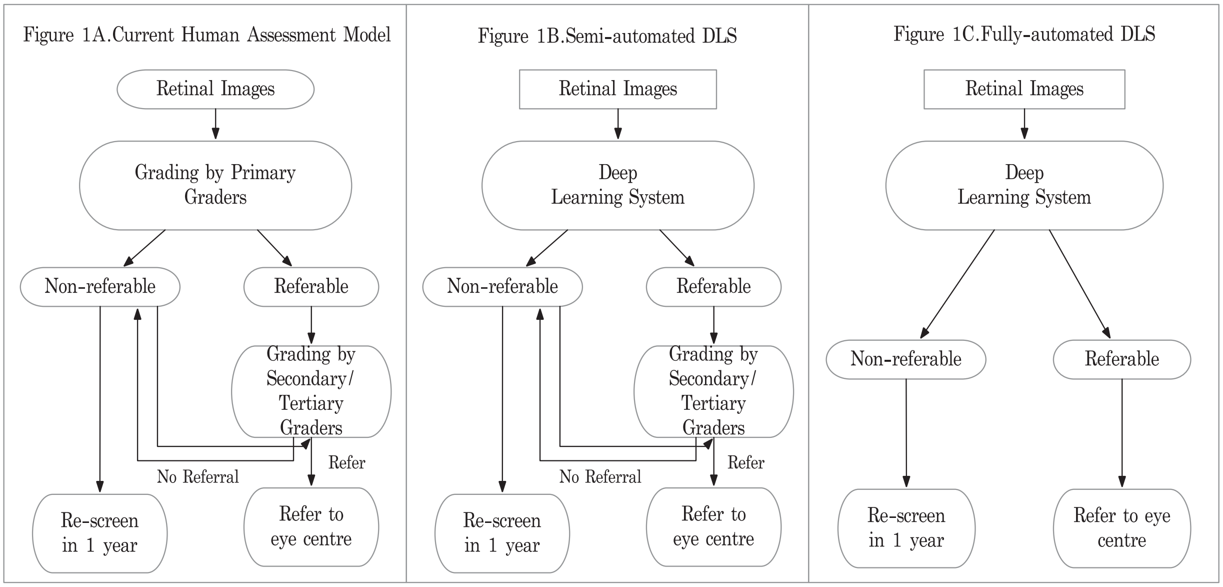

Figure 1. Three potential DR screening models using manual grading (A), semi-automated (B), and fully-automated (C).

Discussion

In this article, we provide a brief summary of the literature regarding the implementation of AI-based DR screening programs, highlighting the need for HEA and safety analyses. A brief discussion on various types of these HEA and safety analysis methods and when to use them in the evaluation of new technologies (e.g., AI) is also included. Practical considerations for the implementation of an AI-based DR screening program have also been outlined, such as the clinical implications of FN rates, FP rates, and image gradeability.

Developing screening tools with a low FN rate has been highlighted as a clinically relevant goal due to patient safety implications. A balance of minimizing both FP and FN needs to be determined based on the intended clinical context. The ideal balance for each context will ultimately be governed by the cost of provider manpower (for adjudication or review of FPs), availability of relevant resources (e.g., various forms of imaging), and the needs of the population it serves (e.g., disease prevalence). Besides screening thresholds, image gradeability is another consideration in evaluating performance of a screening program. This is an important modifiable factor that could affect the HEA of a DLS screening program due to costs involved to reacquire or regrade images and should be included in the evaluation of AI-based solutions. Where relevant resources and technical capabilities are available, additional sources of information, such as three-dimensional OCT scans, may be incorporated to reduce the FP rate in the application of DLS for DR screening, in the same way they have been applied to other eye diseases.[78,79]

This article primarily discusses the role of AI-based solutions for DR screening, which has established cost-effectiveness and has been incorporated in evidence-based practice given improved outcomes with early detection and treatment.[6] AI-based solutions to screen for other major eye diseases, such as glaucoma and age-related macular degeneration (AMD), have also been developed.[80–82] However population screening for these conditions are not yet widely accepted due to inconclusive evidence based on HEA,[83] and clinically acceptable screening performance may vary for these conditions. That being said, incorporation of AI-based solutions may lower manpower costs and help make population screening for these conditions more affordable. Furthermore, Ting et al. have demonstrated that a single AI-based solution for DR screening could be trained to simultaneously detect referable AMD and glaucoma for broad-based eye screening.[16] These considerations will need to be addressed in future research studying the implementation of AI-based solutions for eye screening.

Looking ahead, future research using the tools outlined for HEA and safety analyses are needed to achieve a better understanding of the implementation of AI-based solutions in different settings (e.g., resource-limited settings, remote areas) and with novel screening models (e.g., fully-automated DLS). The required transitions in service delivery along with their associated requirements/costs also need to be investigated. These include transitioning from opportunistic/population DR screening, with or without teleophthalmology services, over to DR screening incorporating AI-based solutions.

Conclusions

To facilitate the real-world integration of AI-based solutions, future studies should also assess the technical feasibility and patient acceptability of implementing these solutions in various primary eye care settings.[85] As these AI-based solutions will influence the practice of ophthalmology and medicine in the near future, it is important to create mechanisms for the direct users (such as optometrists or clinicians) to evaluate and utilize such "black box" AI-based screening programs in clinical practice. Therefore studies to evaluate the health professionals' acceptance and interpretability of AI will be useful to identify barriers to adoption to develop targeted solutions accordingly.[10,13]

Acknowledgments

Supported by the Wellcome Trust [213632/Z/18/Z] (CM).

Disclosure: **Y. Xie**, None; **D.V. Gunasekeran**, None; **K. Balaskas**, None; **P.A. Keane**, None; **D.A. Sim**, None; **L.M. Bachmann**, None; **C. Macrae**, None; **D.S.W. Ting**, Deep learning system for retinal diseases (P)

* YX and DVG contributed equally to this work.

References

1. GBD 2015. Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016;388:1545–1602.
2. GBD 2015. Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016;388:1459–1544.
3. Ting DS, Cheung GC, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol*. 2016;44:260–277.
4. Nangia V, Jonas JB, George R, et al. Prevalence and causes of blindness and vision impairment: magnitude, temporal trends and projections in South and Central Asia. *Br J Ophthalmol*. 2019;103:871–877.
5. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care*. 2004;27:1047–1053.
6. Mohamed Q, Gillies MC, Wong TY. Management of diabetic retinopathy: a systematic review. *JAMA*. 2007;298:902–916.
7. Klein R, Klein BE, Moss SE, Cruickshanks KJ. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. XV. The long-term incidence of macular edema. *Ophthalmology*. 1995;102:7–16.
8. Ferris FL, Patz A. Macular edema. A complication of diabetic retinopathy. *Surv Ophthalmol*. 1984;28(suppl):452–461.
9. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *JAMA*. 2016;316:2366–2367.
10. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167–175.
11. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*, 2019;1:271–297.
12. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28:231–237.
13. Ting DS, Gunasekeran DV, Wickham L, Wong TY. Next generation telemedicine platforms to screen and triage. *Br J Ophthalmol*. 2020;104:299–300.
14. Wilkinson CP, Ferris FL, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110:1677–1682.
15. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131:351–357.
16. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211–2223.

17. Bellemo V, Lim G, Rim TH, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diab Rep*. 2019;19:72.
18. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1: 39.
19. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–1350.
20. Tufail A, Kapetanakis VV, Salas-Vega S, et al. An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness. *Health Technol Assess*. 2016;20:1–72.
21. Raumviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;2:25.
22. Gulshan V, Rajan RP, Widner K, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol*. 2019;137:987–993.
23. Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health*. 2019;1:e35–e44.
24. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health*. 2018;3:e000798.
25. Bjørvig S, Johansen MA, Fossen K. An economic analysis of screening for diabetic retinopathy. *J Telemed Telecare*. 2002;8:32–35.
26. Aoki N, Dunn K, Fukui T, Beck JR, Schull WJ, Li HK. Cost-effectiveness analysis of telemedicine to evaluate diabetic retinopathy in a prison population. *Diabetes Care*. 2004;27:1095–1101.
27. Kirkizlar E, Serban N, Sisson JA, Swann JL, Barnes CS, Williams MD. Evaluation of telemedicine for screening of diabetic retinopathy in the Veterans Health Administration. *Ophthalmology*. 2013;120:2604–2610.
28. Whited JD, Datta SK, Aiello LM, et al. A modeled economic analysis of a digital tele-ophthalmology system as used by three federal health care agencies for detecting proliferative diabetic retinopathy. *Telemed J E Health*. 2005;11:641–651.
29. Li Z, Wu C, Olayiwola JN, Hilaire DS, Huang JJ. Telemedicine-based digital retinal imaging vs standard ophthalmologic evaluation for the assessment of diabetic retinopathy. *Conn Med*. 2012;76:85–90.
30. Kumar S, Tay-Kearney ML, Chaves F, Constable IJ, Yogesan K. Remote ophthalmology services: cost comparison of telemedicine and alternative service delivery options. *J Telemed Telecare*. 2006;12:19–22.
31. Nguyen HV, Tan GS, Tapp RJ, et al. Cost-effectiveness of a National Telemedicine Diabetic Retinopathy Screening Program in Singapore. *Ophthalmology*. 2016;123:2571–2580.
32. Peng J, Zou H, Wang W, et al. Implementation and first-year screening results of an ocular telehealth system for diabetic retinopathy in China. *BMC Health Serv Res*. 2011;11:250.
33. Rachapelle S, Legood R, Alavi Y, et al. The cost-utility of telemedicine to screen for diabetic retinopathy in India. *Ophthalmology*. 2013;120:566–573.
34. Scotland GS, McNamee P, Philip S, et al. Cost-effectiveness of implementing automated grading within the national screening programme for diabetic retinopathy in Scotland. *Br J Ophthalmol*. 2007;91:1518–1523.
35. Philip S, Fleming AD, Goatman KA, et al. The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme. *Br J Ophthalmol*. 2007;91:1512–1517.
36. Scotland GS, McNamee P, Fleming AD, et al. Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *Br J Ophthalmol*. 2010;94:712–719.
37. Prescott G, Sharp P, Goatman K, et al. Improving the cost-effectiveness of photographic screening for diabetic macular oedema: a prospective, multicentre, UK study. *Br J Ophthalmol*. 2014;98:1042–1049.
38. Brown MM, Brown GC, Sharma S, Landy J. Health care economic analyses and value-based medicine. *Surv Ophthalmol*. 2003;48:204–223.
39. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. *BMJ*. 1996;313:275–283.
40. Jakubiak-Lasocka J, Jakubczyk M. Cost-effectiveness versus cost-utility analyses: what are the motives behind using each and how do their results differ?-A Polish example. *Value Health Reg Issues*. 2014;4:66–74.

41. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-Effectiveness in Health and Medicine*. New York, NY: Oxford University Press; 1996. Available at: https://books.google.com.sg/books?id=dazBueIX9L8C . Accessed January 19, 2020.
42. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. Available at: https://www.ncbi.nlm.nih.gov/books/NBK395867/pdf/Bookshelf_NBK395867.pdf . Accessed 20 Dec 2019..
43. Khan T, Bertram MY, Jina R, Mash B, Levitt N, Hofman K. Preventing diabetes blindness: cost effectiveness of a screening programme using digital non-mydriatic fundus photography for diabetic retinopathy in a primary health care setting in South Africa. *Diabetes Res Clin Pract*. 2013;101:170–176.
44. Robinson R. Costs and cost-minimisation analysis. *BMJ*. 1993;307:726–728.
45. Briggs AH, O'Brien BJ. The death of cost-minimization analysis? *Health Econ*. 2001;10:179–184.
46. Ramsey SD, Willke RJ, Glick H, et al. Cost-effectiveness analysis alongside clinical trials II-An ISPOR Good Research Practices Task Force report. *Value Health*. 2015;18:161–172.
47. Dakin H, Wordsworth S. Cost-minimisation analysis versus cost-effectiveness analysis, revisited. *Health Econ*. 2013;22:22–34.
48. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes*. 4th ed. Oxford:Oxford: Oxford University Press; 2015:445. Available at: https://pure.york.ac.uk/portal/en/publications/methods-for-the-economic-evaluation-of-health-careprogrammes(8f69bcee-cdac-44fa-871c-f821470df60a).html. Accessed January 19, 2020.
49. Cartwright WS. Cost-benefit analysis of drug treatment services: review of the literature. *J Ment Health Policy Econ*. 2000;3:11–26.
50. Xie Y, Nguyen DQ, Bellemo V, et al. Cost-effectiveness analysis of an artificial intelligence-assisted deep learning system implemented in the national tele-medicine diabetic retinopathy screening in Singapore. *Invest Ophthalmol Vis Sci*. 2019;60:5471.
51. Petticrew MP, Sowden AJ, Lister-Sharp D, Wright K. False-negative results in screening programmes: systematic review of impact and implications. *Health Technol Assess*. 2000;4:1–120.
52. Nathan DM, Bebu I, Hainsworth D, et al. Frequency of evidence-based screening for retinopathy in type 1 diabetes. *N Engl J Med*. 2017;376:1507–1516.
53. Focal photocoagulation treatment of diabetic macular edema. Relationship of treatment effect to fluorescein angiographic and other retinal characteristics at baseline: ETDRS report no. 19. Early Treatment Diabetic Retinopathy Study Research Group. *Arch Ophthalmol*. 1995;113:1144–1155.
54. Wong RL, Tsang CW, Wong DS, et al. Are we making good use of our public resources? The false-positive rate of screening by fundus photography for diabetic macular oedema. *Hong Kong Med J*. 2017;23:356–364.
55. Petticrew M, Sowden A, Lister-Sharp D. False-negative results in screening programs. Medical, psychological, and other implications. *Int J Technol Assess Health Care*. 2001;17:164–170.
56. Kapetanakis VV, Rudnicka AR, Liew G, et al. A study of whether automated diabetic retinopathy image assessment could replace manual grading steps in the English National Screening Programme. *J Med Screen*. 2015;22:112–118.
57. Henderson C, Knapp M, Fernández JL, et al. Cost effectiveness of telehealth for patients with long term conditions (Whole Systems Demonstrator telehealth questionnaire study): nested economic evaluation in a pragmatic, cluster randomised controlled trial. *BMJ*. 2013;346:f1035.
58. Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. *BMJ*. 2008;336:1281–1283.
59. Byford S, Sefton T. Economic evaluation of complex health and social care interventions. *Natl Inst Econ Rev*. 2003;186:98–108.
60. Anderson R. Systematic reviews of economic evaluations: utility or futility? *Health Econ*. 2010;19:350–364.
61. Deidda M, Geue C, Kreif N, Dundas R, McIntosh E. A framework for conducting economic evaluations alongside natural experiments. *Soc Sci Med*. 2019;220:353–361.
62. Macrae C. Governing the safety of artificial intelligence in healthcare. *BMJ Qual Saf*. 2019;28:495–498.
63. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. 2017;24:1052–1061.
64. Tarpey K, Schaaf E, Lakhani U, Balcitis J. A proactive risk avoidance system using failure mode and effects analysis for "same-name" physician orders. *Jt Comm J Qual Patient Saf*. 2010;36:461–467.

65. Robinson DL, Heigham M, Clark J. Using failure mode and effects analysis for safe administration of chemotherapy to hospitalized children with cancer. *Jt Comm J Qual Patient Saf*. 2006;32:161–166.
66. Coles G, Fuller B, Nordquist K, Weissenberger S, Anderson L, DuBois B. Three kinds of proactive risk analyses for health care. *Jt Comm J Qual Patient Saf*. 2010;36:365–375.
67. Ashley L, Armitage G, Neary M, Hollingsworth G. A practical guide to failure mode and effects analysis in health care: making the most of the team and its meetings. *Jt Comm J Qual Patient Saf*. 2010;36:351–358.
68. Faiella G, Parand A, Franklin BD, et al. Expanding healthcare failure mode and effect analysis: a composite proactive risk analysis approach. *Reliab Eng Syst Saf*. 2018;169:117–126.
69. Leveson N, Thomas J. *STPA Handbook*. United States of America (USA):Massachusetts Institute of Technology (MIT). 2018:3, http://psas.scripts.mit.edu/home/materials/.
70. Salmon PM, Cornelissen M, Trotter MJ. Systems-based accident analysis methods: a comparison of Accimap, HFACS, and STAMP. *Saf Sci*. 2012;50:1158–1170.
71. Chatzimichailidou MM, Ward J, Horberry T, Clarkson PJ. A comparison of the bow-tie and STAMP approaches to reduce the risk of surgical instrument retention. *Risk Anal*. 2018;38:978–990.
72. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125:1264–1272.
73. McLeod RW, Bowie P. Bowtie analysis as a prospective risk assessment technique in primary healthcare. *Policy Pract Heal Saf*. 2018;16:177–193.
74. Kerckhoffs MC, van der Sluijs AF, Binnekade JM, Dongelmans DA. Improving patient safety in the ICU by prospective identification of missing safety barriers using the bow-tie prospective risk analysis model. *J Patient Saf*. 2013;9:154–159.
75. Reason JT. Managing the Risks of Organizational Accidents. Brookfield, VT: Ashgate; 1997. Available at: https://books.google.com/books?id=ZnhRAAAAMAAJ . Accessed January 19, 2020.
76. Cheung CY, Tang F, Ting DSW, Tan GSW, Wong TY. Artificial intelligence in diabetic eye disease screening. *Asia Pac J Ophthalmol (Phila)*. 2019;8:158–164.
77. He J, Baxter SL, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25:30–36.
78. Yoo TK, Choi JY, Seo JG, Ramasubramanian B, Selvaperumal S, Kim DW. The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. *Med Biol Eng Comput*. 2019;57:677–687.
79. Medeiros FA, Jammal AA, Thompson AC. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology*. 2019;126:513–521.
80. Zhen Y, Wang L, Liu H, Zhang J, Pu J. Performance assessment of the deep learning technologies in grading glaucoma severity. 2018:arXiv:1810.13376.
81. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol*. 2019;137:1353–1360.
82. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135:1170–1176.
83. Burr JM, Mowatt G, Hernández R, et al. The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation. *Health Technol Assess*. 2007;11:iii–iv, ix–x, 1–190.
84. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
85. Keel S, Lee PY, Scheetz J, et al. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. *Sci Rep*. 2018;8:4330.