Review Article

Essentials of a Robust Deep Learning System for Diabetic Retinopathy Screening: A Systematic Literature Review

Aan Chu,1 David Squirrell,2 Andelka M. Phillips,3,4 and Ehsan Vaghefi1

1School of Optometry and Vision Science, The University of Auckland, Auckland, New Zealand
2Auckland District Health Board, Auckland, New Zealand
3Te Piringa Faculty of Law, University of Waikato, Hamilton, New Zealand
4HeLEX Centre, Faculty of Law, University of Oxford, Oxford, UK

Correspondence should be addressed to Ehsan Vaghefi; e.vaghefi@auckland.ac.nz

Received 10 August 2020; Revised 20 September 2020; Accepted 3 November 2020; Published 16 November 2020

Copyright © 2020 Aan Chu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This systematic review was performed to identify the specifics of an optimal diabetic retinopathy deep learning algorithm, by identifying the best exemplar research studies of the field, whilst highlighting potential barriers to clinical implementation of such an algorithm. Searching five electronic databases (Embase, MEDLINE, Scopus, PubMed, and the Cochrane Library) returned 747 unique records on 20 December 2019. Predetermined inclusion and exclusion criteria were applied to the search results, resulting in 15 highest-quality publications. A manual search through the reference lists of relevant review articles found from the database search was conducted, yielding no additional records. A validation dataset of the trained deep learning algorithms was used for creating a set of optimal properties for an ideal diabetic retinopathy classification algorithm. Potential limitations to the clinical implementation of such systems were identified as lack of generalizability, limited screeningscope, and data sovereignty issues. It is concluded that deep learning algorithms in the context of diabetic retinopathy screening have reported impressive results. Despite this, the potential sources of limitations in such systems must be evaluated carefully. An ideal deep learning algorithm should be clinic-, clinician-, and camera-agnostic; complying with the local regulation for data sovereignty, storage, privacy, and reporting; whilst requiring minimum human input.

1. Introduction

By 2045, the global incidence of diabetes is projected to reach 629 million adults, with one-third expected to have diabetic retinopathy (DR) [1]. DR remains the leading cause of acquired vision loss in the working-age population [2] and is the most feared microvascular complication of diabetes. Although the aetiology of DR is multifactorial, chronic hyperglycaemia remains the single most important driver of retinal capillary damage. If untreated, DR can result in irreversible vision loss [3] and represents a considerable burden on both the individual and public health systems [4]. Primary prevention of diabetes focusing on modifiable risk factors, such as obesity and lifestyle, has been shown to reduce the development of diabetes. However, such intervention strategies are intensive and require coordinated support networks [5], as well as control of both blood pressure [6] and blood glucose levels [7, 8]. It is now well accepted that screening for DR reduces the risk of vision loss in individuals with diabetes [9–11] and the most effective DR screening modality has been shown to be mydriatic fundus photography [12]. However, DR classification systems and referral pathways differ according to the respective community guidelines [4], and it is often challenging to identify the efficacy of DR screening alone [13], because even in those areas with well-established DR screening programmes in place, patient attendance remains suboptimal [14, 15]. Moreover, although the efficacy of DR screening is not in dispute, systematic DR screening in developing countries is rare [16]. Furthermore, even in developed countries, the disparity in the number of individuals with diabetes and the infrastructure required to sustain DR screening...
programmes, particularly in underserved regions, is expected to widen.

Artificial intelligence (AI), particularly, deep learning, has been touted as the solution to help automate the process of DR screening [17]. Machine learning (ML) is a branch of artificial intelligence and is defined as the study of computer algorithms that allow computer programs to automatically improve through experience [18]. ML relies on working with small to large datasets by examining and comparing the data to find common patterns. ML uses subsets of data to generate an algorithm that may use novel or different combinations of features and weights that can be derived from first principles [19, 20]. In ML, there are four commonly used learning methods, each useful for solving different tasks: supervised, unsupervised, semisupervised, and reinforcement learning [19, 20]. To maximize the chance of generalizability to the performance of the algorithm on unseen data, the training dataset is usually split into a slightly smaller training dataset and a separate validation dataset. Deep learning algorithms (DLAs) are one methodological family of ML based on, e.g., artificial neural networks (ANNs), deep belief networks, recurrent neural networks, or giving a precise example of a feed-forward ANN [21, 22]. Whilst the idea of DLAs is not new, as their origins can be traced back to 1943 [23], the advent of supercomputers and the availability of big data has led to a resurgence of interest in them.

In the context of DR screening, the aim of the DLA is to perform DR grading of fundus photographs, independently of humans. The process of training a DLA to perform DR grading has been described elsewhere [24], but in brief, it involves training a convolutional network (CNN) on a large dataset of images labelled with the correct DR grade: the "ground truth." The DLA then starts assigning a DR grade to each image, and the result generated is then compared with the ground truth. After every comparison, the DLA modifies the neural networks' parameters to improve and maximize its accuracy. This process is repeated until the DLA has "learnt" to assign the correct DR grade to the images in the training dataset. Once training is complete; the DLA's performance is then tested and validated against a bank of unseen images. As challenging as it is to train a DLA, arguably the critical step is the translation into clinical practice, and to date, only a few DLAs have successfully navigated this final hurdle. The objective of this systematic review is to lay the groundwork for both clinicians and developers to evaluate DLAs and highlight potential barriers to their clinical implementation in the context of DR screening. It also aims to stimulate further discussion of appropriate governance in this context. By applying predetermined selection criteria, we aim to only include high-quality studies from our literature search. We will focus on the limitations of the current studies when discussing the barriers to the clinical translation of DR DLAs, as we believe this to be a significant issue.

2. Methods

Several databases including Embase (1980–2019 Week 50) via Ovid, MEDLINE (Ovid MEDLINE Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE Daily, and Ovid MEDLINE 1946-Dec 20, 2019), Scopus, PubMed, and the Cochrane Library were searched. No restrictions on time period, language, or publication type were applied to the electronic database search. A filter was used only for PubMed results to exclude animal studies. All databases were searched on December 20, 2019. The final combination of search terms that returned relevant and nonrestrictive results was ((deep learning OR DNN OR deep neural network OR CNN OR convolutional neural network OR deep learning algorithm OR DLA OR machine learning OR artificial intelligence) AND (diabetic retinopathy OR retinopathy OR maculopathy OR DMO OR DME OR diabetic macular oedema OR diabetic macular oedema)). Only one relevant published systematic review was found from the electronic database search results [25]. We conducted a manual search through the reference lists of this systematic review [25] and that of six review articles found from the electronic database search [17, 26–30], which did not yield any additional results.

All search results obtained from the electronic databases were exported to RefWorks. A total of 1135 results were found, of which we removed 388 duplicates. After excluding the duplicate results, we applied the predetermined selection criteria to the remaining 747 titles and abstracts, if available (Table 1). As the objective at this point was to be deliberately overinclusive, only the inclusion criteria for population, algorithm type, publication category, and image modality were applied. To this end, studies in which a definite decision could not be reached based solely on the title or abstract were still included. This resulted in the exclusion of 682 titles and the inclusion of 67 titles.

We attempted to retrieve the full text of the 65 studies which met stage 1 of the inclusion criteria. Studies that were not available in their entirety were excluded. The remaining studies were assessed against the full inclusion and exclusion criteria. All nonjournal articles, such as conference abstracts or proceedings, comments, and reviews, were excluded, in addition to articles not related to convolutional neural networks. Studies that were not published in English or had incomplete or insufficient information on training, validation, or outcomes were also excluded. The complete selection criteria can be found in Table 1. Note that the inclusion criterion of >5000 images, as DLA training source, was arbitrarily determined. Large training datasets lead to improved performance [17]. However, the exact number of training images needed is uncertain [31]. We identified 15 studies from the electronic database results which met the full selection criteria (Figure 1). AC conducted the data collection and assessment against the selection criteria. Uncertainties in study inclusion were evaluated through discussion with EV and DS until full consensus was reached.

The potential algorithm limitations of each study, as decided by the information provided for its validation set, were then considered (Table 2). The performance of the trained DLAs was also reported. Due to different target conditions amongst the studies and the complexity of reporting all the results, it was decided to focus on the sensitivity and specificity measures for detecting referable
Table 1: Study selection criteria. Stage 1 included population, algorithm type, publication category, and image modality and was applied to 747 articles. The full selection criteria were used for the remaining 65 articles, resulting in the final selection of 13 articles. These were only used as additional search resources.

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Individuals with diabetes (type 1 or 2)</td>
<td>Individuals without diabetes</td>
</tr>
<tr>
<td>Patients with any DR stage and/or DMO</td>
<td>Other retinal diseases</td>
</tr>
<tr>
<td>Populations with DR and other related eye diseases (if data for populations with DR only is separate)</td>
<td>Manual feature construction</td>
</tr>
<tr>
<td><strong>Algorithm type</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Deep learning systems</td>
<td>Expert systems</td>
</tr>
<tr>
<td>Classification tasks (e.g., grading/screening DR)</td>
<td>Segmentation tasks (e.g., lesion quantification)</td>
</tr>
<tr>
<td>Convolutional neural networks (CNN)</td>
<td>Prediction tasks (e.g., future outcomes/prognosis)</td>
</tr>
<tr>
<td><strong>Publication category</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Peer-reviewed</td>
<td>Conference abstracts/proceedings</td>
</tr>
<tr>
<td>Published</td>
<td>Systematic reviews/meta-analyses</td>
</tr>
<tr>
<td><strong>Image modality</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Any retinal camera type</td>
<td>Grey literature (e.g., statistics on diabetes/DR, white papers, clinical practice guidelines)</td>
</tr>
<tr>
<td>Field of view: 40 to 45°</td>
<td>Images from:</td>
</tr>
<tr>
<td>Retinal colour fundus photographs</td>
<td>Smartphones/mobiles</td>
</tr>
<tr>
<td>OCT</td>
<td>Fluorescein angiography</td>
</tr>
<tr>
<td>Stereoscopic imaging</td>
<td>Wide/ultrawide field fundus photography</td>
</tr>
<tr>
<td><strong>Text availability</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Full text available</td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td><strong>DR classification</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Screening or grading DR</td>
<td>Not screening or grading DR</td>
</tr>
<tr>
<td>DR severity scale</td>
<td>No DR severity scale</td>
</tr>
<tr>
<td><strong>Reference standard</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Determined by human graders</td>
<td>Not determined by human graders</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Sensitivity and specificity measures of DR classification</td>
<td>No sensitivity and specificity measures of DR classification</td>
</tr>
<tr>
<td><strong>Training dataset</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>&gt;5000 images</td>
<td>&lt;5000 images</td>
</tr>
<tr>
<td><strong>Validation dataset</strong></td>
<td><strong>Exclusion criteria</strong></td>
</tr>
<tr>
<td>Total number of images</td>
<td>Includes images used for training</td>
</tr>
</tbody>
</table>

3. Results

Of the 747 unique records obtained from the electronic database search, only 15 studies met the selection criteria for this systematic literature review [24, 32–45] (Table 2). The DLA developed by Gulshan et al. [24] functions as the core of several studies, which was originally trained to perform binary classifications of colour fundus photographs as either rDR or nonreferable DR. Krause et al. [33] and a later Gulshan et al. [32] modified this neural network to make multiway classifications into five DR grades according to the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDR). Krause et al. [33] also made other improvements to the original neural network, such as the use of adjudicated data as part of algorithm development. Gulshan et al. [32] then implemented these modifications in their study. The improvements made by Krause et al. became the final DLA used in another study, which evaluated its performance in the DR screening programme in Thailand [34]. Voets et al. [35] attempted to reproduce the results achieved by Gulshan et al. [24]; however, they used publicly available datasets for algorithm training and validation, instead of private datasets. Three of the included studies focused on IDx-DR [39–41]. IDx-DR is the first AI diagnostic medical device authorised by the Food and Drug Association (FDA). Ting et al. [42] developed a DLA that detected rDR, referable AMD, and possible glaucoma. Large datasets of fundus photographs from the Singapore National Diabetic Retinopathy Screening Program were used for DLA training and validation. A secondary validation was performed on ten additional datasets from multiethnic cohorts. Bellomo et al. [43] trained an additional model and combined this with the DLA developed by Ting et al. [42], but only DR and DMO were considered. To better understand the DLA, attention maps were generated to visualise areas in the fundus photographs that contributed most to the DLA output. Visualisation attention maps were also used by Gargeya and Leng [38] for a DLA that detected no DR or DR of any severity. Li et al. [37] developed a DLA for the detection of vision threatening rDR, and Ramachandran et al. [36] validated a third party DLA for detecting rDR. Rogers et al. [44] created a DLA for rDR screening which was then used in...
conjunction with a handheld camera, which is known for lower image qualities compared to stationary cameras. They showed that whilst DLA performed on a par with previous algorithms developed, some mentioned here, it did not perform at the same level using a portable camera setup. Finally, Bhaskaranand et al. [45] assessed the performance of their DLA under real-world conditions and demonstrated that if designed correctly, DLAs can play a vital role as a clinical assistive technology.

Validation datasets were classified as being publicly available or privately held. Of the included studies, four solely used public datasets [35, 38, 39, 45], whilst eight employed privately acquired fundus photographs for validation and four used a combination of both private and public validation datasets. Several studies used the Messidor-2 or Kaggle (from EyePACs screening centres) public datasets as part of the validation dataset. However, a breakdown of the population demographics was not provided in these datasets and those studies that used the public dataset. Ting et al. [42] and Bellemo et al. [46] included the most comprehensive data on patient demographics, detailing systemic risk factors for the development of DR, such as BMI (body mass index), blood pressure, and creatinine levels, in addition to mean age, sex, and ethnicity.

The number of graders used to determine the reference standard for the validation datasets also varied across the studies, from single to 8 graders. Six studies did not detect DMO as part of the validation process, but studies that did detect DMO commonly used hard exudates within 1 disc diameter of the macula as a surrogate for DMO. Li et al. [37] used this criterion or the presence of hard exudates in the macular region that encompassed at least 50% of the disc area. Ting et al. [42] used a less restrictive criterion of any hard exudates in the posterior pole. In addition to exudates, Abr`amoff et al. [39] used retinal thickening or microaneurysms within one disc diameter of the fovea as indicators of DMO. Only one group developed a DLA which also detected possible glaucoma and referable AMD [42].

Ten studies used more than one camera type to take colour fundus photographs for the validation datasets, whilst the known input image resolutions used amongst the studies were \(299 \times 299\) pixels [24, 32, 35, 37], \(512 \times 512\) pixels [38, 39, 42], and \(779 \times 779\) pixels [33, 34]. The number of
fields refers to how many different areas of centration were obtained of the retina in the fundus photographs. Only two studies acquired three fields in a subset of images used for validation [36, 45].

Automated image quality assessment refers to the automatic determination of whether the fundus photographs taken are of adequate quality for grading by a DLA. This was only undertaken by Abramoff et al. [39]. Verbraak et al. [40] also initiated automated image quality assessment but only after manual assessment of the validation dataset images. Image curation is the removal of poor quality or ungradable images from datasets. Only two studies did not curate the validation datasets [42, 46].

Sensitivity, specificity, and AUC measures of the included DLAs are shown in Table 3. As different target conditions and DR grading scales were used, it is difficult to directly compare the included studies. For example, 11 studies defined rDR as having moderate or worse DR, with some including DMO, whereas others used a more severe definition of rDR as preproliferative or worse DR, DMO, or both. Additionally, one study did not use rDR as a target condition, detecting only the absence or presence of DR.

### 4. Discussion

Despite different methods of DLA development, image datasets, and reference standards, a comparison of the included studies is still valuable as it serves to highlight areas that warrant further investigation and improvement. By considering the characteristics of the validation datasets used in the 15 studies, we have identified a number of the current barriers to the clinical implementation of DLAs. These can be categorised into four broad areas, namely, lack of generalizability, limited scope, data protection, and data sovereignty issues. It should be noted though that none of the studies we reviewed herein mentioned intellectual property or privacy issues in any significant way and it is hoped that this article will encourage further discussion of these issues.

One of the key considerations when reviewing the utility of any DLA is to understand its generalizability, as this will determine whether it is suited to the task that it is intended for. Briefly, the generalizability of DLA can be limited by algorithmic bias or by having a scope that does not serve or only incompletely serves those patients with whom it is used. Algorithmic bias is known to be a significant issue in DLA generalizability and subsequent clinical implementation [22, 47]. One recent example of bias has been identified in AI facial recognition systems, where the error rate of gender misclassification in darker-skinned females was 34.7%, compared to 0.8% in lighter-skinned males [48]. Consequently, in order to understand any bias inherent in a DR screening AI, it will be necessary to review whom the DLA was trained upon. A good AI should have access to a large dataset of relevant images. This should include sufficient examples of each class, diseased/nondiseased, etc. This can be challenging to achieve in medicine, where cases of rare diseases or outcomes are, by definition, rare. Whilst some biases may be obvious, others are more subtle and human bias may, therefore, be inadvertently built into a DLA’s decision making [49]. For example, the majority of DLAs developed to date have relied on either private datasets and/
or used datasets that are dominated by a single ethnicity for their training and validation. The AI thus derived may deliver excellent health outcomes for those in the socioeconomic class or ethnic group that the AI was trained on but will perform less well on all others. Adopting the wrong economic class or ethnic group that the AI was trained on will perform less well on all others. Adopting the wrong economic class or ethnic group that the AI was trained on may therefore worsen, not improve, existing health inequalities. DIversification of training datasets and validation of DLAs using data independent of the training dataset are crucial measures to both reduce and evaluate bias [50]. Thus, uncovering bias requires developers to fully disclose the demographics of those it is trained and validated on. Including direct comparisons of validation results difficult. AUC = area under the receiver operating characteristic curve; CI = confidence interval; DMO = diabetic macular oedema; DR = diabetic retinopathy; ETDRS = Early Treatment Diabetic Retinopathy Study; ICDR = International Clinical Diabetic Retinopathy Disease Severity Scale; NHS = English National Health Service; NZ MoH = New Zealand Ministry of Health.

Table 3: Sensitivity, specificity, and AUC validation results of the 15 included studies. Differences in target condition and grading scales made direct comparisons of validation results difficult. AUC = area under the receiver operating characteristic curve; CI = confidence interval; DMO = diabetic macular oedema; DR = diabetic retinopathy; ETDRS = Early Treatment Diabetic Retinopathy Study; ICDR = International Clinical Diabetic Retinopathy Disease Severity Scale; NHS = English National Health Service; NZ MoH = New Zealand Ministry of Health.

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC (95% CI)</th>
<th>Target condition</th>
<th>DR grading scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gulshan et al. [24]</td>
<td>97.5 (95.8–98.7)</td>
<td>93.4 (92.8–94.0)</td>
<td>—</td>
<td>Moderate or worse DR, referable DMO, or both</td>
<td>ICDR</td>
</tr>
<tr>
<td>Gulshan et al. [32]</td>
<td>92.1 (90.1–93.8)</td>
<td>95.2 (94.2–96.1)</td>
<td>0.980</td>
<td>Moderate or worse DR</td>
<td>ICDR</td>
</tr>
<tr>
<td>Krause et al. [33]</td>
<td>97.1 (93.9–100)</td>
<td>89.3 (89.3–96.8)</td>
<td>0.986</td>
<td>Moderate or worse DR</td>
<td>ICDR</td>
</tr>
<tr>
<td>Ruamviboonsuk et al. [34]</td>
<td>96.8 (93.9–100)</td>
<td>95.6 (98.3–96.8)</td>
<td>0.987</td>
<td>Moderate or worse DR</td>
<td>ICDR</td>
</tr>
<tr>
<td>Voets et al. [35]</td>
<td>90.6 (93.3–98.8)</td>
<td>91.0 (84.7–92.8)</td>
<td>0.951</td>
<td>Moderate or worse DR, DMO, or both</td>
<td>ICDR</td>
</tr>
<tr>
<td>Ramachandran et al. [36]</td>
<td>96.0 (93.3–98.8)</td>
<td>90.0 (84.2–89.4)</td>
<td>0.989</td>
<td>Moderate or worse DR, or exudates in 1 disc diameter of the fovea Preproliferative or worse DR, DMO, or both</td>
<td>ICDR and NZ MoH guidelines</td>
</tr>
<tr>
<td>Li et al. [37]</td>
<td>97.0 (96.5–87.9)</td>
<td>91.4 (92.1–94.9)</td>
<td>0.980</td>
<td>Moderate or worse DR or exudates in 1 disc diameter of the fovea</td>
<td>NHS diabetic eye screening guidelines</td>
</tr>
<tr>
<td>Gargeya et al. [38]</td>
<td>93.0 (91.8–94.4)</td>
<td>87.0 (82.8–92.7)</td>
<td>0.940</td>
<td>No DR or any DR</td>
<td>—</td>
</tr>
<tr>
<td>Abr`amoff et al. [39]</td>
<td>96.8 (93.3–98.8)</td>
<td>87.0 (84.2–89.4)</td>
<td>0.989</td>
<td>Moderate or worse DR, DMO, or both</td>
<td>ICDR</td>
</tr>
<tr>
<td>Verbraak et al. [40]</td>
<td>79.4 (96.6–91.2)</td>
<td>93.8 (89.5–94.3)</td>
<td>—</td>
<td>More than mild DR</td>
<td>ICDR</td>
</tr>
<tr>
<td>Abr`amoff et al. [41]</td>
<td>98.9 (66.5–87.9)</td>
<td>92.2 (92.1–94.9)</td>
<td>0.983</td>
<td>More than mild, or DMO, or both</td>
<td>ETDRS</td>
</tr>
<tr>
<td>Ting et al. [42]</td>
<td>98.9 (97.5–99.6)</td>
<td>92.2 (89.5–94.3)</td>
<td>0.973</td>
<td>Moderate or worse DR, DMO, and/or ungradable image</td>
<td>ICDR</td>
</tr>
<tr>
<td>Bellemo et al. [43]</td>
<td>92.3 (90.1–94.1)</td>
<td>89.0 (87.9–90.3)</td>
<td>0.972 (960–0.978)</td>
<td>Moderate or worse DR, DMO, and/or ungradable image</td>
<td>ICDR</td>
</tr>
<tr>
<td>Rogers et al. [44]</td>
<td>93.4% (95%)</td>
<td>94.2% (95%)</td>
<td>98.5% (95%)</td>
<td>Referable DR (RDR) and proliferative DR (PDR) Scottish DR grading scheme</td>
<td>American academy of ophthalmology</td>
</tr>
<tr>
<td>Bhaskaranand et al. [45]</td>
<td>91.3% (95%)</td>
<td>91.1% (95%)</td>
<td>91.7% (95%)</td>
<td>Severe NPDR, proliferative DR</td>
<td>American academy of ophthalmology</td>
</tr>
</tbody>
</table>
DR screening services have developed a granular grading system for diabetic retinopathy, with varying scales across the world. Almost all studies reviewed in this analysis have simplified the various DR grades by combining them into fewer classifications: mild or nonreferrable DR versus moderate or worse DR or “referrable” DR. However, less granular classifications fail to adequately capture the different risk profiles of DR progression. Of the included studies, only Krause et al. [33] developed and provided results for a DLA, which classified fundus photographs into the five-point ICDR grading scale. A DLA that can grade to a more exacting grading system is valuable, as each DR severity level may indicate different management and monitoring pathways depending on regional guidelines and the population involved [51, 52]. However, granular DR classifications in DLAs are more difficult to achieve because, in many datasets, there is a relative paucity of images with more severe and high-risk DR due to the lower prevalence of these grades amongst people with diabetes undergoing screening [53].

Currently, many of the DLAs reviewed do not include diabetic macular oedema (DMO) as a separate entity. DMO is a significant cause of visual impairment in individuals with diabetes [54], and within a standard DR screening programme, both retinopathy and maculopathy need to be detected and graded [55]. Arguably, a DLA designed to be deployed as a tool to deliver DR screening must, therefore, be trained to grade both, and those DLAs which do not detect DMO as a separate entity may result in under-referral of patients with suboptimal patient outcomes. Finally, many of the DLAs published thus far have been trained to read only a single fovea centred image, with many being exposed to a single manufacturer’s camera system. Currently, most DR screening programs, such as the English and the New Zealand National Diabetic Eye Screening Programmes, require 2 image fundus photographs of two 45 degree fields, one fovea centred and one optic disc centred [55]. A DLA that only analyses single field, fovea centred, fundus photographs would not be implementable in this screening setting.

Until recently, it was considered sufficient to simply publish the results of your AI by way of a receiver operator curve, with no explanation as to how the DLA derived this result. This is a critically important issue because what “all” the AI is doing during training is making associations. It is therefore important to be able to assess whether the associations it is making are correct or even relevant. The lack of transparency as to how an AI comes to its decisions is called the “black box phenomenon,” and arguably, if a DLA cannot be understood, how can one assess its reliability and justify its results to patients? This issue can be addressed by the use of attention maps, which highlight which areas within the image the DLA is focusing on when making its decision. With one or two exceptions, most DLAs published thus far have not published such maps. Given that almost any software-based system can be vulnerable in some way, being able to explain black boxes may also be necessary from a debugging perspective.

On a practical level, one aspect which will limit the scope of DR DLA clinical implementation has been the lack of automated image quality assessment or the need for images to be curated manually before being presented to the DLA. Arguably, if manual image quality assessment by professionals is needed prior to an AI inference, the scope of AI implementation is then limited to health centres and providers with such resources and severely reduces the practical utility of the AI. Furthermore, curating images prior to validation of the DLA will likely artificially improve the sensitivity and specificity measures in the test environment, whilst reducing its subsequent utility in a real-world setting.

Although matters such as ethics and intellectual property rights are beyond the scope of this review, a brief discussion is warranted, as concerns around the intellectual property have already been raised [17, 56]. It is also important to recognise the lack of discussion of intellectual property issues in the studies reviewed herein and the need for future work to fill this gap. For instance, clinicians and developers should consider whether the DLA they are using or the software related to it is patentable. They should also consider who has ownership of the algorithm and who owns patient data. In relation to patient rights with respect to their data and medical records, there may be overlap with data protection law. Clinicians will therefore need to consider how they can ensure that patients are informed about the data held about them. Clinicians should also have systems in place to ensure that patient records are kept up to date. Developers should also consider whether the tool they are developing could be treated as a medical device, and if this is the case, they will also need to comply with the frameworks regulating medical devices.

These considerations become more important as there is new evidence that “graph databases” can offer an even higher level of accuracy in matching patient’s needs and healthcare delivery, by combining many different datasets [57]. Graph databases are a technology that is currently used by social media organizations. It is increasingly believed that data-driven approaches can help reduce the current healthcare expenditure [58]. To minimize redundancy and dependency, healthcare data are typically stored and managed using their “normalized” forms [59]. Those normalized tables are later either restructured or “denormalized” for data analytics [60]. By aggregating these forms, a graph database can handle a wide range of graph queries even with big data, whilst revealing many more hidden data about the patient.

Hence, procedures for obtaining informed consent from patients in light of possible reidentification risks and privacy breaches need to be established [61, 62]. It may be helpful for clinicians and developers working with DLA to refer to other electronic consent studies, such as the Dynamic Consent project [63, 64]. Clinicians and developers working in this space also need guidance on how to ensure compliance with both data protection laws (such as the General Data Protection Regulation (GDPR) [65]) and data security requirements. As many countries are currently in the process of reforming their privacy laws in order to align more with the GDPR, privacy and data protection law are currently in a state of flux, and for those utilising data from several countries, adhering to higher data protection standards to begin with may assist in limiting risks to both patients and
organisations in the event of a data breach. Consideration of creating standards for best practices along with other codes of practice could prove useful tools here. Existing privacy and data protection regulators may be able to contribute to this development. In New Zealand, the Office of the New Zealand Privacy Commissioner has previously developed a Health Information Privacy Code [66], which has recently been updated (Health Information Privacy Code 2020: https://privacy.org.nz/privacy-act-2020/codes-of-practice/hipc2020/).

The issue of establishing where liability lies, when a DLA makes an error resulting in misdiagnosis or poor patient outcomes, must also be addressed [67]. The development of medico-legal governance frameworks should precede the implementation of DLAs, with some suggesting a code of conduct upholding the principles of the Hippocratic Oath [62, 68]. Specifically, in considering the development of legal governance frameworks in line with the literature to date, attention should be paid to the following principles: transparency, trust, justice, fairness, equity, nonmaleficence, beneficence, responsibility, accountability, respect for autonomy, sustainability, dignity, and solidarity [69].

Notably, some of the issues mentioned above may need to be addressed in quite distinct ways depending on where the DLA is being developed and deployed. In countries with indigenous populations and other vulnerable groups, DLA developers and clinicians implementing them will need to take account of the specific issues and concerns of these communities [46, 61, 70]. This may necessitate a more cautious approach that gives greater weight to issues of equity, dignity, and social justice, as well as taking account of Indigenous Data Sovereignty [71]. Essentially, the idea of data sovereignty for indigenous peoples can be viewed as referring “to the proper locus of authority over the management of data about indigenous peoples, their territories and ways of life” [71–74].

As our research is conducted in New Zealand, consideration of the Māori perspective in New Zealand is vital. The Te Mana Rāraunga Māori Data Sovereignty Network has developed a Charter, which researchers could refer to as an example of one indigenous perspective on data sovereignty. According to the Charter “Data is a living taonga [treasure] and is of strategic value to Māori and “Māori data is subject to the rights articulated in the Treaty of Waitangi and the UN’s Declaration on the Rights of Indigenous Peoples, to which Aotearoa New Zealand is a signatory.” [75].

Te Mana Rāraunga’s Principles include authority, relationships, obligations, collective benefit, reciprocity, and guardianship. Applying these principles to DLA screening to Māori would mean that they need to be given a voice in how data relating to their community is used and also how these services are offered to their community.

In the New Zealand context, as well as Te Mana Rāraunga, there has also been previous work with the Māori community to develop guidelines for health research. The Māori Health Committee, which is part of the New Zealand Health Research Council, has developed general guidelines for health research involving Māori [76]. Hudson et al. have also developed guidelines for biobanks that handle Māori samples [77]. Both sets of guidelines could serve as examples of the type of work needed where other clinicians and developers want to work with other indigenous peoples. Beaton et al. [78] also provide useful insight into engaging Māori and taking account of the community’s ethical concerns in medical research. Researchers in diabetic retinopathy should think about how they can include Māori and other indigenous groups in an ongoing dialogue in relation to their participation in screening.

Depending on where developers and clinicians are based, they will therefore need to consider how the use of DLA complies with the relevant data protection laws. There is then a need to consider how best to approach these issues prior to wide-scale clinical implementation of a DLA for DR, with the intention of both avoiding harm and enhancing patient trust. It may be useful to utilise focus groups in this context, a move that could provide insight into patients’ views in this context. Including patients’ voices in this space would also help to minimise harm and ensure that respect for dignity and autonomy are upheld.

### 5. Conclusion

In this systematic review, predetermined selection criteria were applied to include high-quality studies. The validation results of 15 studies were analysed to highlight possible barriers currently hindering DLA implementation. We categorised these under lack of generalizability, limited screening scope, data protection, and data sovereignty issues. We do hope that future work will consider the legal and ethical issues raised by DLA in greater depth. There is also a real need to develop the governance framework for DLA before its widespread deployment.

An ideal DLA for DR screening should be camera-, clinic-, and clinician-agnostic, whilst being validated on the local patient demographics. Furthermore, it should include automatic image quality assessment, capable of using uncurated data for granular grading of retinopathy and maculopathy. Finally, this DLA must comply with the local governing body’s requirements for data sovereignty, storage, privacy, and reporting.

A good AI, then, is one that has been trained and validated on large datasets that represent the population in which it is deployed. It is one that reflects the cultural values of the jurisdiction where it is used and it is one that will not further exacerbate existing health inequalities. Increasingly, leading AI scientists are now of the opinion that “Decisions about people should be made by people; AI should be considered a tool to assist human decision making, not its replacement” [79]. Thus, at least for now, it is arguably best to consider DLAs as clinical decision support tools that will aid clinicians and health providers to achieve the best health outcomes for their patients. As such the most effective use of such systems may be to develop new DR DLAs that have a very high negative predictive value to aid the rapid identification of those patients, who are the vast majority, without the disease. This would leave the greatly unburdened human grading team with the task of only needing to assess the small minority with the disease.
Data Availability

This is a review article, and the data are fully available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References


