Hindawi
Journal of Ophthalmology
Volume 2020, Article ID 8841927, 11 pages
https://doi.org/10.1155/2020/8841927

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 Vaghefi**[1]

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

Academic Editor: Ciro Costagliola

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 screening scope, 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 maximise 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 its 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. ‡This number was arbitrarily determined.

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

diabetic retinopathy (rDR), where available. When results on different operating points for high sensitivity or specificity were provided, results reflecting a high sensitivity operating point were included as this is more relevant for screening purposes. If more than one measure of sensitivity and specificity was available for different validation datasets in a study, the best performance achieved was reported.

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. Bellemo 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

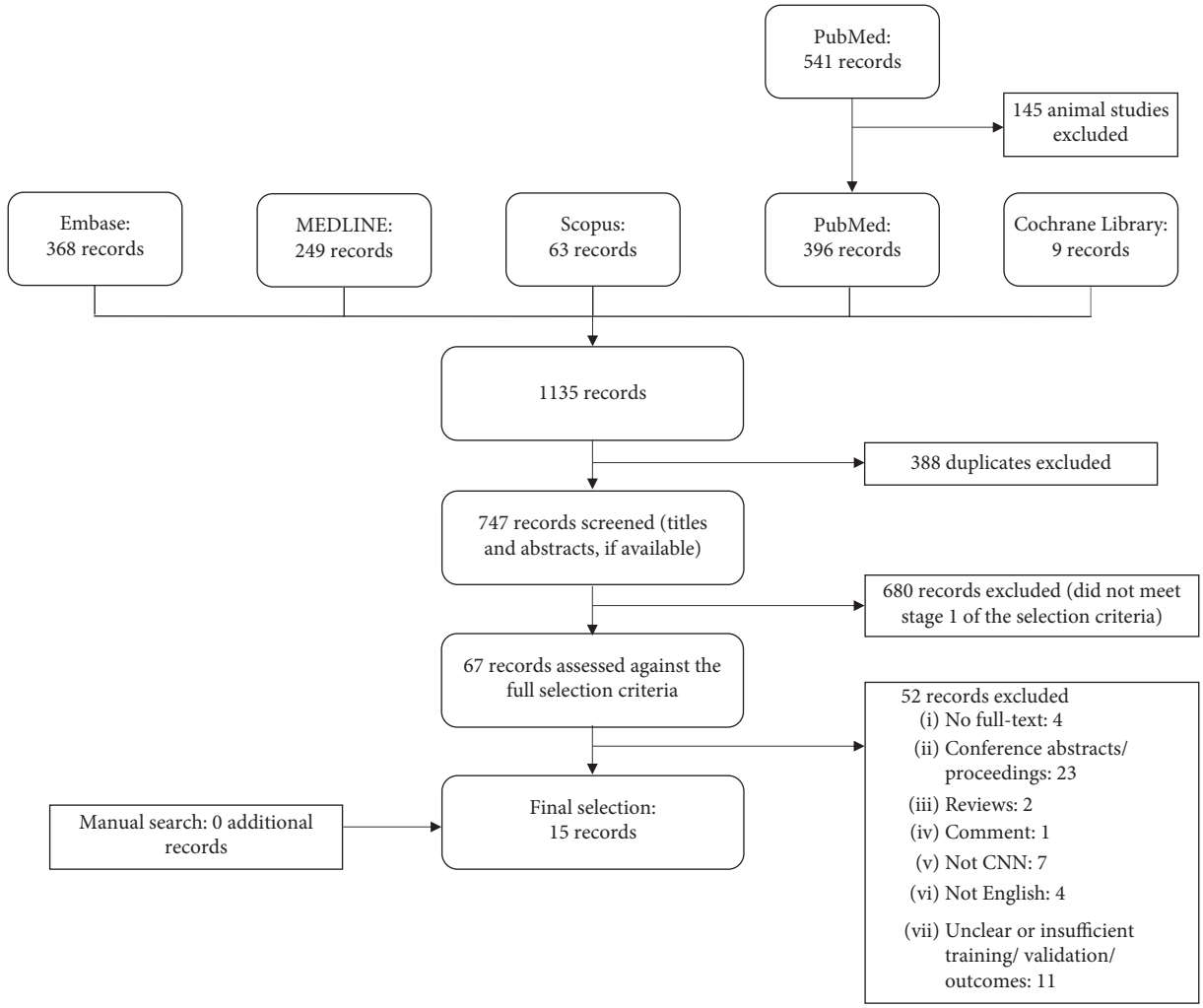

Figure 1: Flowchart of the study selection process. CNN = convolutional neural network.

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, Abràmoff 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 299×299 pixels [24, 32, 35, 37], 512×512 pixels [38, 39, 42], and 779×779 pixels [33, 34]. The number of

Table 2: Potential sources of algorithm limitations present in the validation datasets of the 13 + 2 included studies. ✓ = yes; ✕ = no; † = combination; ‡ = exact number unspecified; § = unknown.

	Public dataset	Reference standard grader number	DMO detection	>1 camera type	Image resolution (pixels)	Number of fields	Automated image quality assessment	Image curation	Other disease detection
Gulshan et al. [24]	†	7-8	✓	✓	299 × 299	1	✕	✓	✕
Gulshan et al. [32]	✕	3 + ‡	✓	✓	299 × 299	1	✕	✓	✕
Krause et al. [33]	✕	6	✓	§	779 × 779	1	✕	✓	✕
Ruamviboonsuk et al. [34]	✕	2–3	✓	✓	779 × 779	1	✕	✓	✕
Voets et al. [35]	✓	‡	✕	✓	299 × 299	1	✕	§	✕
Ramachandran et al. [36]	†	2+‡	✕	✓	§	1–3	✕	✓	✕
Li et al. [37]	†	1–5	§	✓	299 × 299	§	✕	✓	✕
Gargeya et al. [38]	✓	‡	✕	✓	512 × 512	§	✕	✓	✕
Abràmoff et al. [39]	✓	3	✓	✕	§	1	✕	✕	✕
Verbraak et al. [40]	✕	>2	✕	✕	§	2	†	✓	✕
Abràmoff et al. [41]	✕	‡	✕	✕	§	2	✓	✓	✕
Ting et al. [42]	✕	‡	✓	✓	512 × 512	§	✕	✕	✓
Bellemo et al. [43]	✕	2	✓	✕	512 × 512	1	✕	✓	✕
Rogers et al. [44]	†	>2	✓	✓	§	1	✕	✓	✕
Bhaskaranand et al. [45]	✓	>2	✕	✓	§	3	✕	✓	✕

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 Abràmoff 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/

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.

	% (95% CI)				
	Sensitivity	Specificity	AUC (95% CI)	Target condition	DR grading scale
Gulshan et al. [24]	97.5 (95.8–98.7)	93.4 (92.8–94.0)	—	Moderate or worse DR, referable DMO, or both	ICDR
Gulshan et al. [32]	92.1 (90.1–93.8)	95.2 (94.2–96.1)	0.980	Moderate or worse DR	ICDR
Krause et al. [33]	97.1	92.3	0.986	Moderate or worse DR	ICDR
Ruamviboonsuk et al. [34]	96.8 (93.9–100)	95.6 (98.3–96.8)	0.987 (0.977–0.995)	Moderate or worse DR	ICDR
Voets et al. [35]	90.6	84.7	0.951 (0.947–0.956)	Moderate or worse DR	ICDR
Ramachandran et al. [36]	96.0	90.0	0.980 (0.973–0.986)	Moderate or worse DR, or exudates in 1 disc diameter of the fovea	ICDR and NZ MoH guidelines
Li et al. [37]	97.0	91.4	0.989	Preproliferative or worse DR, DMO, or both	NHS diabetic eye screening guidelines
Gargeya et al. [38]	93.0	87.0	0.940	No DR or any DR	—
Abràmoff et al. [39]	96.8 (93.3–98.8)	87.0 (84.2–89.4)	0.980 (0.968–0.992)	Moderate or worse DR, DMO, or both	ICDR
Verbraak et al. [40]	79.4 (66.5–87.9)	93.8 (92.1–94.9)	—	More than mild DR	ICDR
Abràmoff et al. [41]	87.2 (81.8–91.2)	90.7 (88.3–92.7)	—	More than mild, or DMO, or both	ETDRS
Ting et al. [42]	98.9 (97.5–99.6)	92.2 (89.5–94.3)	0.983 (0.972–0.991)	Moderate or worse DR, DMO, and/or ungradable image	ICDR
Bellemo et al. [43]	92.3 (90.1–94.1)	89.0 (87.9–90.3)	0.973 (0.969–0.978)	Moderate or worse DR, DMO, and/or ungradable image	ICDR
Rogers et al. [44]	93.4% (95% CI: 90.8–95.8)	94.2% (95% CI: 91.0–97.2)	98.5% (95% CI: 97.8–99.2)	Referable DR (RDR) and proliferative DR (PDR)	Scottish DR grading scheme
Bhaskaranand et al. [45]	91.3% (95% CI: 90.9%–91.7%)	91.1% (95% CI: 90.9%–91.3%)	0.965 (95% CI: 0.963–0.966)	Severe NPDR, proliferative DR	American academy of ophthalmology

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 AI 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. Publishing the demographics of the training and validation datasets is, therefore, crucial to understanding the generalizability of the DLA. Our review reveals that most of the major studies published thus far have used relatively small private datasets for the validation of the DLAs. Moreover, of these, only two published significant demographic information [42, 46]. Clearly, this needs to be addressed in further studies.

Another bias inherent within any algorithm is the integrity of the underlying "ground truth" and how this was derived. Across the studies included in this review, there was great diversity in the number and experience of the graders used to determine the reference standard of the validation datasets. Additionally, each study followed a different protocol to generate its reference standard. Arguably, when establishing "ground truth," a majority vote may not be sufficiently rigorous. Instead, a live adjudicated consensus of several retinal specialists should be incorporated into future studies involving DLAs to improve algorithm accuracy and, subsequently, patient outcomes. Although live adjudication involves greater resources at the outset of training, only a small proportion of images may need to be subject to this [33]. This was demonstrated by Ruamviboonsuk et al. [34], where expert graders only adjudicated a subset of images that the algorithm and regional graders disagreed on. Further investigation into establishing a method of adjudication that is time and resource-efficient and yields improved algorithm performance is needed.

Defining the scope within which the DLA has been trained and validated is clearly also important, as this will have a direct impact on its generalizability. Most established

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 nonreferable DR versus moderate or worse DR or "referable" 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 underreferral of patients with suboptimal patient outcomes. Finally, many of the DLAs published thus far have been trained to read only a single foveal 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 Raraunga 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 Raraunga'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 Raraunga, 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 in 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

[1] P. Saeedi, I. Petersohn, P. Salpea et al., "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the international diabetes federation diabetes atlas, 9th edition," *Diabetes Research and Clinical Practice*, vol. 157, Article ID 107843, 2019.

[2] E. Fenwick, G. Rees, K. Pesudovs et al., "Social and emotional impact of diabetic retinopathy: a review," *Clinical & Experimental Ophthalmology*, vol. 40, no. 1, pp. 27–38, 2012.

[3] M. W. Stewart, H. W. Flynn, S. G. Schwartz, and I. U. Scott, "Extended duration strategies for the pharmacologic treatment of diabetic retinopathy: current status and future prospects," *Expert Opinion on Drug Delivery*, vol. 13, no. 9, pp. 1277–1287, 2016.

[4] D. S. W. Ting, G. C. M. Cheung, and T. Y. Wong, "Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review," *Clinical & Experimental Ophthalmology*, vol. 44, no. 4, pp. 260–277, 2016.

[5] J. Tuomilehto, J. Lindström, J. G. Eriksson et al., "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance," *New England Journal of Medicine*, vol. 344, no. 18, pp. 1343–1350, 2001.

[6] Group UPDS, "Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: ukpds 38," *BMJ: British Medical Journal*, vol. 317, no. 7160, pp. 703–713, 1998.

[7] Group UPDS, "Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34)," *The Lancet*, vol. 352, no. 9131, pp. 854–865, 1998.

[8] E. Stefánsson, T. Bek, M. Porta, N. Larsen, J. K. Kristinsson, and E. Agardh, "Screening and prevention of diabetic blindness," *Acta Ophthalmologica Scandinavica*, vol. 78, no. 4, pp. 374–385, 2000.

[9] Group UPDS, "Photocoagulation treatment of proliferative diabetic retinopathy: clinical application of diabetic retinopathy study (DRS) findings, DRS report number 8," *Ophthalmology*, vol. 88, no. 7, pp. 583–600, 1981.

[10] Group ETDRSR, "Early photocoagulation for diabetic retinopathy: ETDRS report number 9," *Ophthalmology*, vol. 98, no. 5, pp. 766–785, 1991.

[11] Q. Mohamed, M. C. Gillies, and T. Y. Wong, "Management of diabetic retinopathy: a systematic review," *JAMA*, vol. 298, no. 8, pp. 902–916, 2007.

[12] A. Hutchinson, A. McIntosh, J. Peters et al., "Effectiveness of screening and monitoring tests for diabetic retinopathy—a systematic review," *Diabetic Medicine*, vol. 17, no. 7, pp. 495–506, 2000.

[13] J. H. Vallance, P. J. Wilson, G. P. Leese, R. McAlpine, C. J. MacEwen, and J. D. Ellis, "Diabetic retinopathy: more patients, less laser: a longitudinal population-based study in Tayside, Scotland," *Diabetes Care*, vol. 31, no. 6, pp. 1126–1131, 2008.

[14] K. N. D. Van Eijk, J. W. Blom, J. Gussekloo, B. C. P. Polak, and Y. Groeneveld, "Diabetic retinopathy screening in patients with diabetes mellitus in primary care: incentives and barriers to screening attendance," *Diabetes Research and Clinical Practice*, vol. 96, no. 1, pp. 10–16, 2012.

[15] G. P. Leese, P. Boyle, Z. Feng, A. Emslie-Smith, and J. D. Ellis, "Screening uptake in a well-established diabetic retinopathy screening program: the role of geographical access and deprivation," *Diabetes Care*, vol. 31, no. 11, pp. 2131–2135, 2008.

[16] D. S. Friedman, F. Ali, and N. Kourgialis, "Diabetic retinopathy in the developing world: how to approach identifying and treating underserved populations," *American Journal of Ophthalmology*, vol. 151, no. 2, pp. 192–194, 2011.

[17] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," *Progress in Retinal and Eye Research*, vol. 67, pp. 1–29, 2018.

[18] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning and Data Mining*, Springer, Cham, Switzerland, 2017.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, Cham, Switzerland, 2013.

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media, Cham, Switzerland, 2009.

[21] D. Singh, E. Merdivan, I. Psychoula et al., "Human activity recognition using recurrent neural networks," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction,* Springer, Cham, Switzerland, 2017.

[22] A. Holzinger, "Introduction to machine learning & knowledge extraction (make)," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 1–20, 2019.

[23] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[24] V. Gulshan, L. Peng, M. Coram et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[25] K. B. Nielsen, M. L. Lautrup, J. K. H. Andersen, T. R. Savarimuthu, and J. Grauslund, "Deep learning-based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance," *Ophthalmology Retina*, vol. 3, no. 4, pp. 294–304, 2019.

[26] L. Balyen and T. Peto, "Promising artificial intelligence-machine learning-deep learning algorithms in ophthalmology," *The Asia-Pacific Journal of Ophthalmology*, vol. 8, no. 3, pp. 264–272, 2019.

[27] D. T. Hogarty, D. A. Mackey, and A. W. Hewitt, "Current state and future prospects of artificial intelligence in ophthalmology: a review," *Clinical & Experimental Ophthalmology*, vol. 47, no. 1, pp. 128–139, 2019.

[28] Z. Li, S. Keel, and M. He, "Can artificial intelligence make screening faster, more accurate, and more accessible?" *The Asia-Pacific Journal of Ophthalmology*, vol. 7, no. 6, pp. 436–441, 2018.

[29] E. Rahimy, "Deep learning applications in ophthalmology," *Current Opinion in Ophthalmology*, vol. 29, no. 3, pp. 254–260, 2018.

[30] D. S. W. TingL. R. Pasquale et al., "Artificial intelligence and deep learning in ophthalmology," *British Journal of Ophthalmology*, vol. 103, no. 2, pp. 167–175, 2019.

[31] B. J. Erickson, P. Korfiatis, T. L. Kline, Z. Akkus, K. Philbrick, and A. D. Weston, "Deep learning in radiology: does one size fit all?" *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 521–526, 2018.

[32] V. Gulshan, R. P. Rajan, K. Widner et al., "Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India," *JAMA Ophthalmology*, vol. 137, no. 9, pp. 987–993, 2019.

[33] J. Krause, V. Gulshan, E. Rahimy et al., "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.

[34] P. Ruamviboonsuk, J. Krause, P. Chotcomwongse et al., "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program," *NPJ Digital Medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[35] M. Voets, K. Møllersen, and L. A. Bongo, "Reproduction study using public data of: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *PLoS One*, vol. 14, no. 6, Article ID e0217541, 2019.

[36] N. Ramachandran, S. C. Hong, M. J. Sime, and G. A. Wilson, "Diabetic retinopathy screening using deep neural network," *Clinical & Experimental Ophthalmology*, vol. 46, no. 4, pp. 412–416, 2018.

[37] Z. Li, S. Keel, C. Liu et al., "An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs," *Diabetes Care*, vol. 41, no. 12, pp. 2509–2516, 2018.

[38] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.

[39] M. D. Abràmoff, Y. Lou, A. Erginay et al., "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative Opthalmology & Visual Science*, vol. 57, no. 13, pp. 5200–5206, 2016.

[40] F. D. Verbraak, M. D. Abramoff, G. C. F. Bausch et al., "Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting," *Diabetes Care*, vol. 42, no. 4, pp. 651–656, 2019.

[41] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–8, 2018.

[42] D. S. W. Ting, C. Y.-L. Cheung, G. Lim 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*, vol. 318, no. 22, pp. 2211–2223, 2017.

[43] V. Bellemo, Z. W. Lim, G. Lim et al., "Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study," *The Lancet Digital Health*, vol. 1, no. 1, pp. e35–e44, 2019.

[44] T. W. Rogers, J. Gonzalez-Bueno, F. R. Garcia et al., "Evaluation of an AI system for the detection of diabetic retinopathy from images captured with a handheld portable fundus camera: the MAILOR AI study," *Eye*, 2019.

[45] M. Bhaskaranand, C. Ramachandra, S. Bhat et al., "The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes," *Diabetes Technology & Therapeutics*, vol. 21, no. 11, pp. 635–643, 2019.

[46] V. Bellemo, G. Lim, T. H. Rim et al., "Artificial intelligence screening for diabetic retinopathy: the real-world emerging application," *Current Diabetes Reports*, vol. 19, no. 9, p. 72, 2019.

[47] J. Zou and L. Schiebinger, "AI can be sexist and racist—it's time to make it fair," *Nature Publishing Group*, vol. 559, pp. 324–326, 2018.

[48] J. Buolamwini and T. Gebru, "Gender shades: intersectional accuracy disparities in commercial gender classification," *Fairness, Accountability and Transparency*, vol. 81, pp. 1–15, 2018.

[49] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care—addressing ethical challenges," *New England Journal of Medicine*, vol. 378, no. 11, p. 981, 2018.

[50] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, Sydney, Australia, 2015.

[51] J. W. Y. Yau, S. L. Rogers, R. Kawasaki et al., "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, vol. 35, no. 3, pp. 556–564, 2012.

[52] P. H. Scanlon, "Screening intervals for diabetic retinopathy and implications for care," *Current Diabetes Reports*, vol. 17, no. 10, p. 96, 2017.

[53] R. Sayres, A. Taly, E. Rahimy et al., "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019.

[54] T. A. Ciulla, A. G. Amador, and B. Zinman, "Diabetic retinopathy and diabetic macular edema: pathophysiology, screening, and novel therapies," *Diabetes Care*, vol. 26, no. 9, pp. 2653–2664, 2003.

[55] P. H. Scanlon, "The English national screening programme for diabetic retinopathy 2003–2016," *Acta Diabetologica*, vol. 54, no. 6, pp. 515–525, 2017.

[56] K. A. Mikk, H. A. Sleeper, and E. Topol, "Patient data ownership-reply," *JAMA*, vol. 319, no. 9, pp. 935-936, 2018.

[57] Y. Park, M. Shankar, B.-H. Park, and J. Ghosh, "Graph databases for large-scale healthcare systems: a framework for efficient data management and data services," in *Proceedings of the 2014 IEEE 30th International Conference on Data Engineering Workshops*, Chicago, IL, USA, March 2014.

[58] C. Meier, "A role for data: an observation on empowering stakeholders," *American Journal of Preventive Medicine*, vol. 44, no. 1, pp. S5–S11, 2013.

[59] A. Silberschatz, H. F. Korth, and S. Sudarshan, *Database System Concepts*, McGraw-Hill, New York, NY, USA, 1997.

[60] J. Ravi, Z. Yu, and W. Shi, "A survey on dynamic web content generation and delivery techniques," *Journal of Network and Computer Applications*, vol. 32, no. 5, pp. 943–960, 2009.

[61] Nature, "Time to discuss consent in digital-data studies," *Nature*, vol. 572, no. 7767, p. 5, 2019.

[62] P. Balthazar, P. Harri, A. Prater, and N. M. Safdar, "Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 580–586, 2018.

[63] J. Kaye, E. A. Whitley, D. Lund, M. Morrison, H. Teare, and K. Melham, "Dynamic consent: a patient interface for twenty-first century research networks," *European Journal of Human Genetics*, vol. 23, no. 2, pp. 141–146, 2015.

[64] F. K. Dankar, M. Gergely, B. Malin et al., "Dynamic-informed consent: a potential solution for ethical dilemmas in

population sequencing initiatives," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 913–921, 2020.
[65] P. Regulation, *General Data Protection Regulation*, Intouch, Shenzhen, China, 2018.
[66] P. Commissioner, *Health Information Privacy Code 1994*, KB Print Ltd., Auckland, New Zealand, 1994.
[67] F. Pesapane, C. Volonté, M. Codari, and F. Sardanelli, "Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States," *Insights Into Imaging*, vol. 9, no. 5, pp. 745–753, 2018.
[68] D. Helbing, B. S. Frey, G. Gigerenzer et al., *Will Democracy Survive Big Data and Artificial Intelligence? in: Towards Digital Enlightenment*, pp. 73–98, Springer, Cham, Switzerland, 2019.
[69] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.
[70] J. Blumenstock, "Don't forget people in the use of big data for development," *Nature*, vol. 561, no. 7722, pp. 170–172, 2018.
[71] A. Phillips, *Buying Your Self on the Internet. Wrap Contracts and Personal Genomics*, Edinburgh University Press, Edinburgh, UK, 2019.
[72] R. James, R. Tsosie, R. Tsosie et al., "Exploring pathways to trust: a tribal perspective on data sharing," *Genetics in Medicine*, vol. 16, no. 11, pp. 820–826, 2014.
[73] P. Hummel, M. Braun, S. Augsberg, and P. Dabrock, "Sovereignty and data sharing," *ITU Journal: ICT Discoveries*, vol. 25, no. 8, 2018.
[74] A. Ballantyne and R. Style, "Health data research in New Zealand: updating the ethical governance framework," *NewZealand Medical Journal*, vol. 130, no. 1464, pp. 64–71, 2017.
[75] M. D. S. Network, "Te mana raraunga—māori data sovereignty network charter," 2016.
[76] Health Research Council of New Zealand, *Guidelines for Researchers on Health Research Involving Maori 2010*, Health Research Council of New Zealand, Auckland, New Zealand, 2010.
[77] M. Hudson, A. Beaton, M. Milne et al., *He Tangata Kei Tua: Guidelines for Biobanking with Māori. Māori and Indigenous Governance Centre*, University of Waikato, Hamilton, New Zealand, 2016.
[78] A. Beaton, M. Hudson, M. Milne et al., "Engaging Māori in biobanking and genomic research: a model for biobanks to guide culturally informed governance, operational, and community engagement activities," *Genetics in Medicine*, vol. 19, no. 3, pp. 345–351, 2017.
[79] R. Galloway, "Improving crop yields with mobile phones," 2019, https://podcasts.apple.com/zw/podcast/improving-crop-yields-with-mobile-phones/id73331490?i=1000459902946.