Current Practices in Ophthalmology

Series Editor: Parul Ichhpujani

Parul Ichhpujani
Sahil Thakur *Editors*

Artificial Intelligence and Ophthalmology

Perks, Perils and Pitfalls

Springer

Current Practices in Ophthalmology

Series Editor

Parul Ichhpujani
Department of Ophthalmology
Government Medical College and Hospital
Chandigarh, India

This series of highly organized and uniform handbooks aims to cover the latest clinically relevant developments in ophthalmology. In the wake of rapidly evolving innovations in the field of basic research, pharmacology, surgical techniques and imaging devices for the management of ophthalmic disorders, it is extremely important to invest in books that help you stay updated. These handbooks are designed to bridge the gap between journals and standard texts providing reviews on advances that are now part of mainstream clinical practice. Meant for residents, fellows-in-training, generalist ophthalmologists and specialists alike, each volume under this series covers current perspectives on relevant topics and meets the CME requirements as a go-to reference guide. Supervised and reviewed by a subject expert, chapters in each volume provide leading-edge information most relevant and useful for clinical ophthalmologists. This series is also useful for residents and fellows training in various subspecialties of ophthalmology, who can read these books while at work or during emergency duties. Additionally, these handbooks can aid in preparing for clinical case discussions at various forums and examinations.

More information about this series at http://www.springer.com/series/15743

Parul Ichhpujani • Sahil Thakur
Editors

Artificial Intelligence and Ophthalmology

Perks, Perils and Pitfalls

Springer

Editors
Parul Ichhpujani
Department of Ophthalmology
Government Medical College and Hospital
Chandigarh
India

Sahil Thakur
Ocular Epidemiology Research Group
Singapore Eye Research Institute
Singapore

ISSN 2523-3807 ISSN 2523-3815 (electronic)
Current Practices in Ophthalmology
ISBN 978-981-16-0633-5 ISBN 978-981-16-0634-2 (eBook)
https://doi.org/10.1007/978-981-16-0634-2

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

About the Authors

Parul Ichhpujani is currently a professor in the Department of Ophthalmology at Government Medical College and Hospital, Chandigarh, India, where she is chiefly responsible for glaucoma and neuro-ophthalmology services. She completed her glaucoma training at the Advanced Eye Centre, Postgraduate Institute of Medical Education and Research, Chandigarh, India, and in a subsequent clinical research fellowship, under Dr. George L Spaeth, at Wills Eye Institute, Philadelphia, USA. She currently serves on the education committee of the World Glaucoma Association and is the associate managing editor of the *Journal of Current Glaucoma Practice*. She was ranked among the Powerlist 2015 for the "Best 40 ophthalmologists under 40."

An avid researcher, Dr. Ichhpujani has coauthored three books: *Pearls in Glaucoma Therapy*, *Living with Glaucoma*, and *Smart Resources in Ophthalmology*; and has edited another eight: *Expert Techniques in Ophthalmology*, *Glaucoma: Basic and Clinical Perspectives*, *Manual of Glaucoma*, *Clinical Cases in Glaucoma: An Evidence-Based Approach*, *Glaucoma: Intraocular Pressure and Aqueous Dynamics*, *Current Advances in Ophthalmic Technology*, *Glaucoma*, and *Ophthalmic Instruments and Surgical Tools*. She is also the Springer series editor for the "Current Practices in Ophthalmology" series. She has also contributed several research articles and book chapters in national and international books and serves as a reviewer for many ophthalmology journals. She is heavily involved in resident and fellow teaching.

Sahil Thakur, MBBS, MS graduated from Government Medical College and Hospital, Chandigarh, India. He is currently a clinical research fellow at the Singapore Eye Research Institute. Apart from pursuing clinical ophthalmology, he has a keen interest in developing affordable medical technology and clinical photography. He is also involved in the research and development of digital solutions for faster and efficient diagnosis of common ophthalmic disorders. Previously he has coauthored a book, *Smart Resources in Ophthalmology: Applications and Social Networking*, and has more than 50 peer-reviewed publications. Dr. Thakur also has an interest in developing tools for ophthalmic education and has published two Android applications (Ophthalmaster and Ophthalminion) for use by residents and medical school students.

A Clinician's Introduction to Artificial Intelligence

1

Sahil Thakur and Ching-Yu Cheng

1.1 Artificial Intelligence

To understand the concept of artificial intelligence (AI) and how it is being used in applications today, we first need to understand the concept of intelligence. The term intelligence is derived from the Latin noun '*intellēctus*' or verb '*intelligere*', which means to comprehend or perceive. This concept is however abstract and is better understood with examples of different types of intelligence and how humans display them.

1. Visual-spatial: physical environment characteristics (architects when designing a building according to terrain and surroundings, navigating a boat in water).
2. Kinaesthetic: body movements (technical skill and precision of a ballerina, surgeons or athletes).
3. Creative: novel thought, typically expressed in art, music and writing (imagination-driven authors, painters and musicians).
4. Interpersonal: interaction with others (interviewers, shopkeepers, businessmen).
5. Intrapersonal: self-realisation (meditation, goal planning, self-preservation).

S. Thakur (✉)
Ocular Epidemiology Research Group, Singapore Eye Research Institute, Singapore

C.-Y. Cheng
Ocular Epidemiology Research Group, Singapore Eye Research Institute, Singapore, Singapore

Ophthalmology and Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore, Singapore

Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore

P. Ichhpujani, S. Thakur (eds.), *Artificial Intelligence and Ophthalmology*, Current Practices in Ophthalmology,
https://doi.org/10.1007/978-981-16-0634-2_1

6. Linguistic: manipulation of words for communication (day-to-day communication).
7. Logical-mathematical: calculations, identifying patterns, analysing relationships (logic, puzzles, computing numbers).

From this classification, it is easy to understand what AI today is capable of and where we may be heading in the future. The simulation potential of logical-mathematical intelligence is the maximum and early development in AI almost exclusively focused on this domain. Robotics aims to mimic the kinaesthetic intelligence while sensor-driven (LiDAR scanners in self driven cars) applications leverage on visual-spatial intelligence. Chatbots are trying to mimic linguistic and interpersonal intelligence while the creative and intrapersonal intelligence are domains with limited to no simulation potential. Utility of algorithms has been explored to create music and draw art, but this is mainly driven by logical-mathematical intelligence.

When we understand what an algorithm can and cannot do, that is when we can maximise the utility of the algorithm. Thus, it is imperative to stay away from over-optimistic predictions and avoid false promises to increase the acceptability of algorithms and their potential widespread use. Some algorithms that have achieved this level of acceptance are the 'search engine' algorithms that offer personalised search results, spam filters in email clients, recommendations in applications like Netflix or Amazon and computational photography algorithms on mobile devices. Algorithms that have been developed for use in hospitals however are yet to see such levels of acceptance. This has been due to the inherent nature of patient–physician relationship, potential regulatory hurdles and multiple types of bias that confound these algorithms. However, with FDA approvals being given to 64 algorithms (SaMD: software as medical device) over the last 3 years, we can expect widespread availability of these options for clinicians in the future [1, 2]. Currently for ophthalmology only the IDx-DR has been approved as an autonomous AI diagnostic system for diabetic retinopathy [3–5].

1.2 The Past and What We Can Learn from It

The earliest examples of humans trying to build intelligent devices were the abacus like devices namely the *nepohualtzintzin* (Aztecs), *suanpan* (Chinese) or the *soroban* (Japan) [6]. These devices though based on simple concepts, reduced the time required for mathematical computations. This concept of reducing time and effort for repetitive computational tasks remains one of the driving concepts behind algorithm development.

The *Antikythera* mechanism was another ancient computing device that was probably used to track dates of important events, predict eclipses and even planetary motions [7]. Ramon Llull's *Ars Magna* was another device that used simple paper-based rotating concentric circle to generate combinations of new words and ideas. It was a rudimentary step towards generating a logical system to produce knowledge

[8]. Examples of such systems also exist in fictional literature like the book-writing engine in the city of Lagado in Gulliver's Travels. Attempts to create a similar algorithm include the RACTER program which generated text for the first computer authored book titled 'The Policeman's Beard is Half Constructed' in 1983 [9].

Perhaps one of the most significant inventions in primitive computing was the Difference Engine, proposed by Charles Babbage in 1822 [10]. In addition to this engine, Babbage also wanted to create the Analytical Engine which could be programmed using punch cards and had separate areas for number storage and computation. Ada Lovelace, the daughter of English poet, Lord Byron, gave the specifications for designing a program for this Engine. She is now considered by many as the first computer programmer [11].

Currently we know that data is stored in computers as a series of binary 1 s and 0 s called bits. Eight bits make up one byte. The fundamentals of this concept were published in a book, titled, 'An Investigation into the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities', by George Boole in 1854 [12]. He wanted to reduce logic to simple algebra involving only 0 and 1, with three simple operations: and, or and not. Boolean algebra, which is named after him, is one of the foundations of this digital age.

Over the next few decades, there were incremental improvements in algorithms for applications like optical character recognition (OCR), handwriting recognition (HWR) and speech synthesis. The next breakthrough was the 1943 paper 'A Logical Calculus of the Ideas Immanent in Nervous Activity' by Warren McCulloch and Walter Pitts [13]. In this paper, they described the basic mathematical model of the biological neuron. This formed the basis for the development of artificial neural networks (ANN) and deep learning (DL).

ENIAC, short for Electronic Numerical Integrator and Computer, was unveiled in 1946 and represented the pinnacle of specialised electronic, reprogrammable, digital computers built to solve a range of computing problems [14]. This started the race for development of powerful computer hardware for specialised operations by different countries. However, by today's standards even the Apollo Space Mission Guidance Computer (AGC) only had 64 KB memory and operated at 0.043 MHz, when compared to today's smartphones running with GHz speed processors (A14 chips in iPhones and iPads run at 3.0GHz and thus clock 70,000 times faster) shows how far we have come in terms of computing power due to the development of semiconductor technology [15, 16].

The term, 'artificial intelligence' (AI) was coined by John McCarthy at the Dartmouth conference for experts in this field in 1956 [17]. The expectations from this conference were extremely high despite limited computing power and hardware at that time. Inability to meet the hype generated by this conference, thus led to the AI winters of 1974–1980 and 1987–1993 [18].

Meanwhile during this time interesting developments were taking place in the backdrop, like:

- Rosenblatt concept of the perceptron [19] (1957).
- Arthur Lee Samuel's concept of machine learning [20] (1959).

- ELIZA: The program that could respond to text input simulating a conversation [21] (1964).
- Early deep learning using supervised multilayer perceptrons (1965).
- MYCIN: Rule-based expert system to identify sepsis and to recommend antibiotics [22] (1970).
- Fuzzy logic and its applications in automation [23] (1965–1974).
- Lighthill Report (criticised the utter failure of artificial intelligence in achieving its 'grandiose objectives') that triggered the first AI winter [24] (1973).
- Joseph Weizenbaum's early idea of ethics in AI, suggestion that AI should not be used as substitutes for humans in jobs requiring compassion, interpersonal respect, love, empathy and care [25] (1976).
- Expert system boom driven by LISP machines; however, LISP was soon overtaken by IBM/Apple with more powerful and cheaper consumer desktop computers, this led to collapse of the demand for expert systems [26] (1980–1987).
- Alex Waibel's Time Delay Neural Network (TDNN) which was the first convolutional network [27] (1987).
- Moravec's paradox: Tasks simple for humans like walking, talking, face/voice recognition are difficult for AI while humanly complex computational tasks involving mathematics and logic are simple [28] (1988).
- Yan LeCun developed system to recognise handwritten ZIP codes [29] (1989).
- Chinook (checkers playing algorithm) vs Marion Tinsley [30] (1994).
- IBM Deep Blue (chess playing algorithm) vs Garry Kasparov [31] (1997).
- Logistello (othello playing algorithm) vs Takeshi Murakami [32] (1997).
- Oh and Jung demonstrated power of graphical processing units (GPUs) for network training [33] (2004).
- ImageNet database [34] (2009).
- IBM DeepQA-based Watson winning the quiz show Jeopardy [35] (2011).
- Google DeepMind AlphaGo (based on ANN and Monte Carlo tree search algorithm defecting Lee Sedol) and AlphaGo Zero (trained by self-play without using previous data) which subsequently defected AlphaGo [36] (2017).
- Adversarial patches and perturbations [37, 38] (2018).
- Stanford death predictor [39] (2019).

Perhaps the most important developments that renewed interest in the field of AI and allowed widespread access over the last decade are the availability of large amounts of data and increased computational power at cheaper costs using modalities like graphical processing units (GPUs). ImageNet has especially been used to train popular models like the AlexNet [40], VGG16 [41], Inception modules [42] and the currently used ResNet [43].

Other datasets are also available for applications like music, facial recognition, text and speech processing [44]. As AI is a rapidly evolving field, today new innovations also happen with the same pace. However, understanding the history of AI is vital in predicting how it may affect the future. In further sections, we discuss why AI has become so popular today and how it may help in optimising patient care by evolving into an effective decision support system.

1.3 Why Should a Clinician Bother About AI?

A quick PubMed search shows how the number of articles published in the field of AI has grown to 112,594 results with 35,140 (31.2%) being published since 2018 [45]. Another insight comes from the Gartner Hype Index that monitors and predicts how a technology will evolve over time [46]. Machine learning (ML) was at the peak of inflated expectation indicating impact of publicity and expectations in 2016, DL at the same peak in 2018. These peaks also translate to the increase in applications that were developed using these technologies in this time. PubMed search shows a total of 49, 721 results till 2020 for ML, with 3885 results in 2016, 5217 in 2017 and 8169 in 2018. The last 2 years have seen 24,230 results which is 48.73% of total results [47]. Similarly, for DL, PubMed search shows 18,082 results till 2020 with 3020 results in 2018, 5401 in 2019 and 7383 in 2020 [48]. The last 2 years represent 70.7% of the total results. These numbers show how these technologies are being increasingly tried and tested for use in medicine.

Due to the lack of special training for understanding or evaluating these applications or their underlying concepts, a lot of effort has been recently initiated to make the clinicians more aware and sensitised about the use of AI in providing patient care [49–51]. In the next section, we describe a checklist approach to reading an AI paper with emphasis on evidence assessment and evaluation of future potential for translation to clinical use. We believe that this approach can help in better understanding of the scientific merit of the publication and its potential impact on care delivery practice patterns.

1.4 How to Read an Artificial Intelligence Paper?

Jaeschke et al. provided a framework to evaluate diagnostic tests in clinical medicine [52]. We have expanded the same framework to include relevant information about AI-based algorithms. We will initially describe the framework and then provide example of using the framework [53, 54]. The framework is as follows:

- Step 1: Evaluate if the study results are valid.
  Primary Guide
  – Was there an independent, blind comparison with a reference standard?
  – Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?
    For AI-based algorithms these can be adapted as:
      Are the datasets appropriate and described in sufficient detail?
      Was the gold standard for algorithm training appropriate and reliable?

Secondary Guide

- Did the results of the test being evaluated influence the decision to perform the reference standard?
- Were the methods for performing the test described in sufficient detail to permit replication?

  For AI-based algorithms these can be adapted as:

    Is the methodology of algorithm development described in sufficient detail to allow replication?

    Are the algorithm/datasets used available for external validation?

- Step 2: Evaluate the presented results.
  - Are likelihood ratios for the test results presented or data necessary for their calculation provided?

    For AI-based algorithms these can be adapted as:

      Are adequate and appropriate performance metrics reported? [50].
- Step 3: Evaluate the utility of results in providing care for your patients.
  - Will the reproducibility of the test result and its interpretation be satisfactory in my setting?
  - Are the results applicable to my patient?
  - Will the results change my management?
  - Will the patients be better off because of the test?

    For AI-based algorithms these can be adapted as:

      Are the findings of the algorithm explainable? Does the algorithm exhibit generalisability (can it be easily adapted for a different machine input or population)? Was the original algorithm performance too optimistic?

      Has the algorithm been validated in my local population?

      Is there any independent comparison of the algorithm with existing standard of care? Is there a cost-effectiveness analysis for rationale of algorithm use?

      Will there be a significant impact on patient well-being after algorithm deployment? Is there an attempt to measure this impact?

Table 1.1 shows how this framework can be used to evaluate an artificial intelligence paper.

Table 1.1 Framework for evaluation of artificial intelligence papers in medicine (adopted from Jaeschke et al.) [52]

Paper Title: Clinically applicable deep learning for diagnosis and referral in retinal disease [53]. *Purpose: Develop an artificial intelligence-based patient triage system using 3D OCT data*		
Step 1: Evaluate if the study results are valid		
• Was there an independent, blind comparison with a reference standard?	Are the datasets appropriate and described in sufficient detail?	The authors describe in detail the training set for OCT (Topcon) segmentation (877 scans), validation set for segmentation (224 scans), training set for classification (14,884 scans), validation set for classification (993 scans) and the testing set for comparison of algorithm (997 random scans) with standard of care
• Did the patient sample include an appropriate spectrum of patients to whom the diagnostic test will be applied in clinical practice?	Was the reference standard for algorithm training/testing appropriate and reliable?	The data for training the segmentation algorithm was manually segmented by trained ophthalmologists, reviewed and edited by senior ophthalmologists. The training set for classification used labels from automatic note search and trained ophthalmologists/optometrists reviewed the scans. The validation set for classification was graded by three junior graders, while for the test set, the referral gold standard was from full patient clinical records to determine the diagnosis and referral path considering subsequently obtained information. The algorithm performance was compared to four medical retina consultant ophthalmologists and four specialist optometrists
• Did the results of the test being evaluated influence the decision to perform the reference standard?	Did the results of the algorithm influence the decision to perform the reference standard?	No, the referral gold standard was retrospective data based on full clinical records of patients undergoing current standard of care
• Were the methods for performing the test described in sufficient detail to permit replication?	Is the methodology of algorithm development described in sufficient detail to allow replication? Are the algorithm/datasets used available for external validation?	The authors describe the algorithm (U-net architecture) in detail but mention that the data is not available in the public domain and may be available on request subject to local and national ethical approvals. In a subsequent paper, the authors mention about releasing the segmentation algorithm and dataset in the public domain for validation [54]

(continued)

Table 1.1 (continued)

Step 2: Evaluate the presented results		
• Are likelihood ratios for the test results presented or data necessary for their calculation provided?	Are adequate and appropriate performance metrics reported? [50]	The authors report ROC curves, confusion matrices, total error rates and impact of additional information (OCT alone, OCT + fundus + full case summary) on expert referral decisions. The algorithm had an AUC of 99.21 and error rate of 5.5% (55/997)
Step 3: Evaluate the utility of results in providing care for your patients		
• Will the reproducibility of the test result and its interpretation be satisfactory in my setting?	Are the findings of the algorithm explainable? Does the algorithm exhibit generalisability? Was the original algorithm performance too optimistic?	The authors report data for generalising using of the algorithm using another OCT device. (Spectralis), though initially the algorithm performs poorly and has error rate of 46.6% for referral decisions, retraining of the segmentation algorithm improves the AUC to 99.93 and reduces error rate to 3.4% (4/116). This shows that the algorithm is flexible and adaptable to a different machine. The authors also report results in a third OCT machine (Cirrus 5000) where initial error rate of 16.4% was reduced to 9.8% after retraining the segmentation algorithm The developers of the algorithm also tried to incorporate elements of explainable artificial intelligence by providing a segmentation maps with highlighted retinal structure, pathology, artefacts and predicted diagnostic probabilities and referral suggestions. However, in the videos provided as supplementary material, the automatic segmentation is not always accurate
• Are the results applicable to my patient?	Has the algorithm been validated in my local population?	No, the results are from the patient population at Moorfields eye hospital, London, United Kingdom. It will need further validation in different ethnic populations and research settings before it can be applicable to your patients
• Will the results change my management?	Is there any independent comparison of the algorithm with existing standard of care? Is there a cost effectiveness analysis for rationale of algorithm use?	The algorithm was compared to 4 medical retina consultant ophthalmologists and 4 specialist optometrists. The algorithm performed as well or outperformed the experts. There was no attempt however to assess the cost effectiveness of the algorithm as compared to standard of care
• Will the patients be better off because of the test?	Will there be a significant impact on patient well-being after algorithm deployment? Is there an attempt to measure this impact?	The algorithm has potential to be deployed as a clinician decision support tool but immediate impact on patient well-being cannot be assessed. No attempt was made by the authors to measure this impact in the real world