A MIXED METHODS EVALUATION OF ARTIFICIAL INTELLIGENCE-ENABLED MACULA SERVICES

by

JEFFRY HOGG

A thesis submitted to the University of Newcastle upon Tyne for the degree of DOCTOR OF PHILOSOPHY

Population Health Science Institute
Faculty of Medical Sciences
University of Newcastle upon Tyne

January 2024

Abstract

Introduction

Ophthalmology incurs an increasing number of NHS hospital outpatient appointments, more than any other specialty. Neovascular age-related macular degeneration (nAMD) makes the third largest contribution to ophthalmology appointments. A sight-threatening imbalance between demand and capacity for these macular appointments could be addressed by a well-validated artificial intelligence (AI) technology, yet to be prospectively applied in NHS research or practice.

This thesis aims to explore the factors which limit clinical AI implementation and develop actionable solutions for the implementation of AI-enabled macula services in the NHS.

Methods

The thesis applies a pragmatist approach, draws on the disciplinary field of implementation science, and uses mixed methods. Qualitative evidence synthesis, qualitative interviewing, a retrospective diagnostic accuracy study and theoretically informed analyses are performed.

Findings

Five distinct stakeholder groups illuminate the interdependent factors that influence clinical AI implementation. AI-enabled macula services offer broad value recognised by most stakeholders who prioritise evidence that implementation will not lead to sight loss. A simulated AI-enabled medical device used a candidate AI technology to independently make nAMD treatment decisions with less undertreatment and less overtreatment than consultant-led-care. The AI-enabled intervention to operationalise this medical device should delegate treatment planning decisions away from ophthalmologists and partly apply freed resources to improve patient-clinician communication quality. Healthcare pathway analysis proposed AI use and training to optimise the safety, effectiveness, and fairness of AI-enabled macula services.

Conclusions

The novelty of clinical AI and limited connectivity between its stakeholders sustain the implementation gap observed generally and within macula services. The problem of mismatched demand and capacity in macula services is real for all key stakeholders and an AI solution appears able to offer value to each. NHS organisations are free to locally implement AI-enabled macula services and this thesis provides evidence to inform if and how they choose to proceed.

Word count (excluding reference sections and appendices): 55,497

Acknowledgements

Compiling this thesis has provided ample opportunity to reflect on the personal journey it represents for me. Thinking back to 2018, developing the ideas underlying the initial proposal, I could not have anticipated the distance and dimensions through which the journey would carry me. It has brought me professional flexibility when it was invaluable for my family and I, awareness of the broader channels through which I can pursue my goals and connections with many brilliant and inspiring individuals.

It is not possible or reasonable to comprehensively list and thank those individuals here, but the fundamental contributions of my four supervisors cannot be passed by: professors Katie Brittain (Newcastle University), Pearse Keane (University College London), Gregory Maniatopoulos (Leicester University) and Dawn Teare (Newcastle University). Having such varied, expert, and dependable guidance has been the key enabler to the production of this thesis. I would particularly like to thank each of you for believing in me when I exercised my tendency to optimism and for patiently illuminating the occasions where that optimism got the better of me. I have learnt so much from each of you about not just what you do, but why and how to do it. You have allowed me to enthusiastically rebuff every misplaced grimace I receive upon disclosing my supervisory headcount.

So much so, that I and this thesis came to depend on an even wider source of tuition and guidance for which I owe deep thanks for. To Fiona Beyer for her expert tuition and oversight of the evidence synthesis foundation to this work. To James Talks for imparting his deep and authentic insights into the clinical and operational problem which we aimed to address. To Trevor and Janet Lunn for generously sharing their lived experiences with neovascular age-related macular degeneration and guiding the alignment of the research with those experiences. To Rashmi Kumar, Rosemary Nicholls, Angela Quinn and Christine Sinnett for their patience a comradery in refining the evidence we created to enhance its public service. I am also grateful to the National Institute for Health and Care Research and UK taxpayer for funding this thesis and my experience in its entirety, and for making the provision of such opportunities a priority for myself and others.

At the risk of cliché, the submission of this thesis feels much more of a beginning than an end, and I can only thank those that have enabled it and attempt to contribute to that service for others. Only time will tell if this sense of beginning is another exercise in optimism or good intuition. Let’s see.

Contents

List of Figures

List of Tables

List of Abbreviations

AFC	Agenda For Change
AI	Artificial Intelligence
AIaMD	Artificial Intelligence as a Medical Device
AMD	Age-related macular degeneration
CDS	Clinical Decision Support
CI	Confidence Interval
CLC	Consultant-Led-Care
DL	Deep Learning
EMR	Electronic Medical Record
ENTREQ	Enhancing Transparency in Reporting the Synthesis of Qualitative research
ERIC	Expert Recommendations for Implementing Change
ETDRS	Early Treatment of Diabetic Retinopathy Study
F2F	Face-to-Face
FDA	Food and Drug Administration
FITT	Fit between Individual, Task and Technology
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GP	General Practitioner
HMW	How Might We
ICB	Integrated Care Board
IDAOPI	Income Deprivation Affecting Older People Index
IQR	Interquartile Range
IRF	Intraretinal Fluid
IT	Information Technology

IUS	Intended Use Statement
IVT	Intravitreal Treatment
JBI	Joanna Briggs Institute
LR	Likelihood Ratio
MAA	Medical Algorithmic Audit
MHRA	Medicines and Healthcare products Regulatory Agency
ML	Machine Learning
MRC	Moorfields Reading Centre
nAMD	Neovascular Age-related Macular Degeneration
NASSS	Non-adoption, Abandonment, Scale-up, Spread and Sustainability
NB	Notified Body
NHS	National Health Service
NICE	National Institute for health and Care Excellence
NPV	Negative Predictive Value
NS	Not Specified
NuTH	Newcastle upon Tyne Hospitals
OCT	Optical Coherence Tomography
OECD	Organization for Economic Cooperation and Development
PACS	Picture Archiving and Communication System
PESTLE	Political, Economic, Sociological, Technological, Legal and Environmental
PPV	Positive Predictive Value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRN	Pro Re Nata
QMS	Quality Management System
RETREAT	Review question-Epistemology-Time or Timescale-Resources-Expertise-Audience and Purpose-Type of data
rNPV	Relative Negative Predictive Value
rPPV	Relative Positive Predictive Value

SaMD	Software as a Medical Device
SAS	Specialty doctors and Associate Specialists
SHRM	Subretinal Hyperreflective Material
SRF	Subretinal Fluid
ST	Specialty Training year
T&E	Treat and Extend
T2DM	Type 2 Diabetes Mellitus
TAM	Technology Acceptance Model
TMF	Theories, Models or Framework
TN	True Negative
TP	True Positive
UK	United Kingdom
VA	Visual Acuity
VEGF	Vascular Endothelial Growth Factor

Chapter 1: Introduction

This introductory chapter aims to equip readers with the necessary foundations in clinical context, technology, and the disciplinary field of implementation science.

Clinical context: Ophthalmology is the busiest hospital outpatient specialty, with just 24% of departments reporting adequate consultant capacity to meet clinical demand in 2022. Around 17% of all these ophthalmology outpatient appointments relate to the management of age-related macular degeneration, only the neovascular form (nAMD) of which is treatable with 3-8 intravitreal injections (IVI) per year, often for more than a decade. Once a diagnosis of nAMD has been made, the major demand for clinician input comes from repeated decisions about when sequential IVIs should be planned for each patient. These decisions are largely made from optical coherence tomography (OCT) imaging of the macula. If the requirement for clinician involvement in this single repeated decision-making process could be reduced, then there could be significant relief to the sight-threatening imbalance between capacity and demand in UK Ophthalmology.

Technology: Conceived more than 70 years ago, artificial intelligence (AI) is receiving increasing attention and investment from government, industry and academia in healthcare and other sectors. This most recent revival of interest in the field came from a high-profile step change in the performance of a subtype of AI technologies in 2012, known as deep learning. Despite the optimism for deep learning enabled healthcare interventions, there are few examples of their real-world implementation and even fewer with clear evidence of positive outcomes. This persistent translational gap has become known as the "AI chasm". As a clinical specialty that is relatively dependent on image interpretation, particularly in the management of nAMD, ophthalmology is a priority specialty for AI innovation. However, there are no known investigations of AI-enabled macula services.

Implementation science: Implementation science is a relatively young disciplinary field drawing on diverse epistemological approaches and disciplines across a spectrum of research and practice. Its pragmatic goal of bridging know-do gaps to improve real-world healthcare necessitates this multi-disciplinary approach. Implementation research focuses on shaping and evaluating healthcare interventions and implementation strategies. A key aspect of implementation science is the application of theories, models or frameworks (TMF) to inform or interpret implementation efforts and outcomes in a particular healthcare context.

Aim: This thesis aims to explore the factors which sustain the AI chasm in healthcare generally and develop actionable solutions for the implementation of AI-enabled macula services in the National Health Service (NHS).

1.1 Service Demands on UK Ophthalmology

A threat to the efficacy of United Kingdom (UK) National Health Service (NHS) ophthalmology services is posed by insufficient ophthalmologist availability, in the face of a growing clinical need .[1] A 2022 workforce census found that only 25% of hospital eye units report an adequate consultant workforce for current service demands, and 25% of consultant ophthalmologists plan to leave the NHS in the next 5 years. Strategic adaptions to meet the national shortfall have included a small increase to speciality training capacity, extending the roles of allied health professionals and revising accreditation pathways besides specialty training.[2-4] The impact of these interventions so far appears inadequate as ophthalmology continue to incur more clinic appointments, having overtaken orthopaedics as the busiest hospital outpatient speciality in 2017/18 (Figure 1).[5] In the short to mid-term, this capacity-demand mismatch has been exacerbated by the suspension of many ophthalmology services during Covid-19, producing lists of over 600,000 awaiting hospital ophthalmology appointments.[6, 7] Ordered by size, the three largest disease contributors to ophthalmology service demands are cataract, glaucoma and age-related macular degeneration (AMD). All these disproportionately affect older populations who were encouraged to defer non-urgent care episodes due to their greater risk from Covid-19.[5] In the long-term, the context of an ageing population also means that demand for ophthalmology services can be expected to continue rising and so sustainable solutions are required to meet these needs.[8]

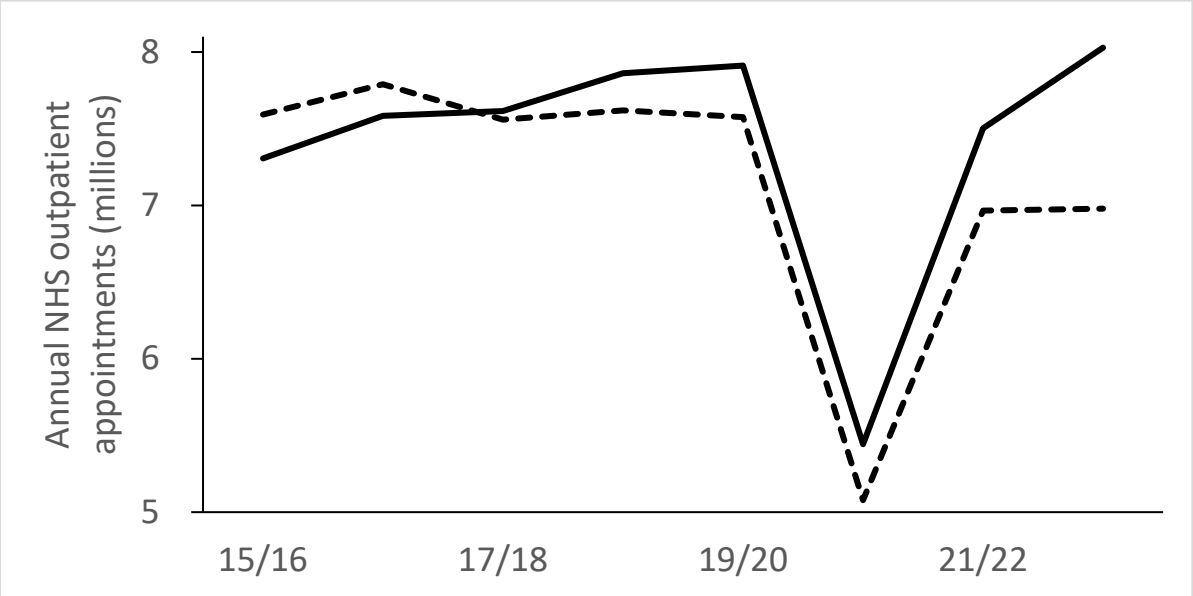


Figure 1. Outpatient appointments from NHS England in ophthalmology (full line) and orthopaedics (dashed line).[5]

1.1.1 Age-related macular degeneration

The clinical coding method to establish disease-linked clinic demands within the NHS is retrospective and performed by non-clinical staff, but the resulting data suggest a contribution of 17.0% of all ophthalmology outpatient appointments in 2022/23.[5] Consequently, AMD services present a potentially impactful exemplar in which to rebalance clinical capacity and demand. AMD is the leading cause for certification of visual impairment in the UK and affects just over a quarter of Europeans aged over 60 to some extent.[9, 10] AMD affects the macula, which is the part of the retina which facilitates high detail central vision and which structures toward the front of the eye focus light toward (Figure 2).

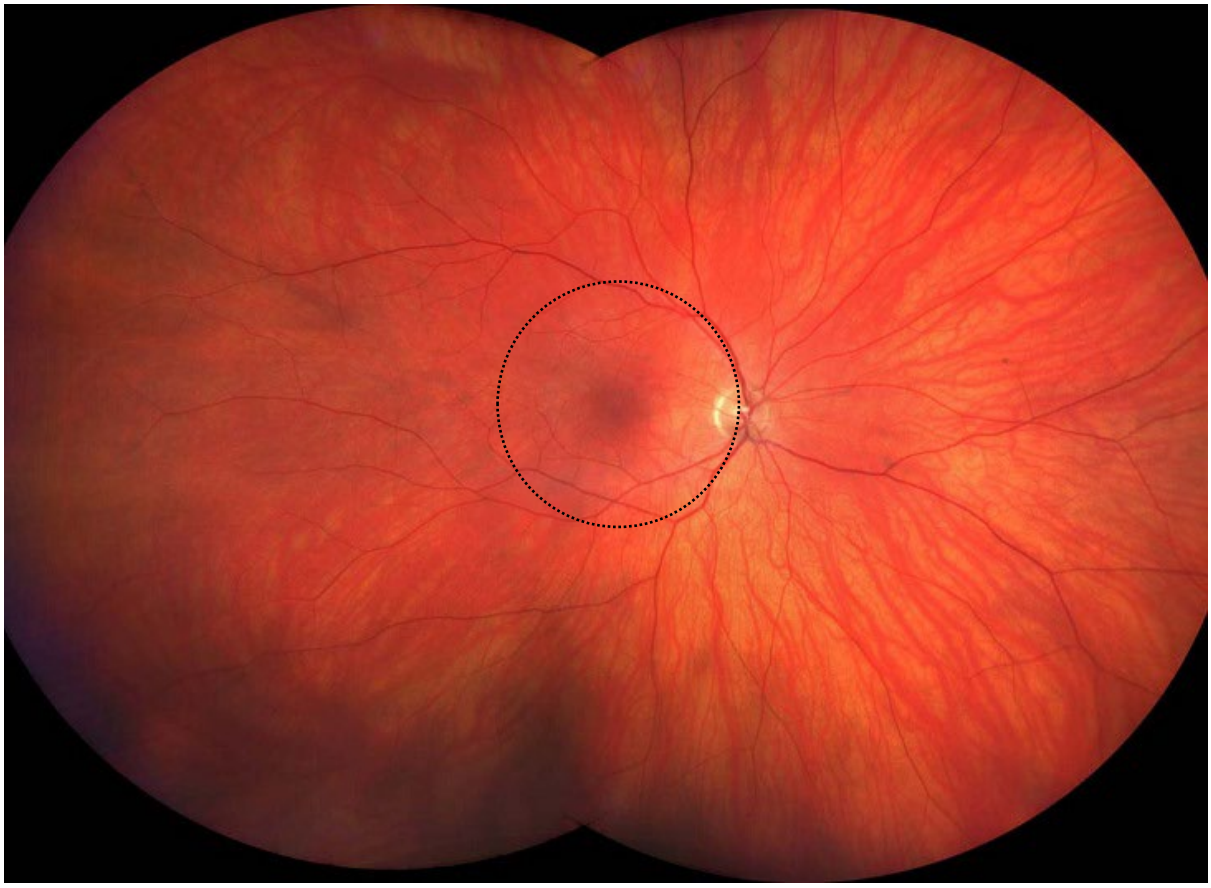


Figure 2. Ultra-wide field enface photograph of a healthy right retina with the macula marked by a dotted black circle.

Like all the retina, the macula converts light that enters the eye into electrical impulses which travel to the visual cortex in the brain and are experienced as vision. AMD is a progressive disease, but it often affects people towards the end of their lives and progresses slowly. Consequently, most people with AMD die from unrelated causes before their symptoms pose any substantial limit on their quality of life. In terms of the appearance of the retina, all stages are characterised by drusen. Drusen are lipid deposits within the basement membrane of the retina that appear as yellow-coloured dots of varying size (Figure 3).

Figure 3. Right macula (greatly magnified compared to figure 2) with large drusen and small spots of pigmentary change centrally; an example of intermediate age-related macular degeneration.

Whilst early stages of AMD are the most prevalent, the more severe stages of the disease can remove the central vision of one or both eyes completely. A classification system of early, intermediate and late is used to characterise individuals' disease and also the likelihood of them progressing to late AMD in the near future (Table 1).[11]

Table 1. Clinical classification of age-related macular degeneration (AMD) proposed by the Beckman Initiative for Macular Research Classification Committee.[11]

AMD stage	Normal ageing changes	Early AMD	Intermediate AMD	Late AMD	
				Neovascular	Geographic atrophy
Clinical feature	Small drusen (<63 µm)	Medium drusen (63-125 µm)	Medium drusen and pigmentary change or	Intra or sub-retinal fluid +/- haemorrhage	Patchy atrophy of the retinal pigment epithelium

			large drusen (>125 µm)		and outer retina

Unfortunately, there are no well-evidenced treatments for most stages of the disease. Management often focuses on lifestyle advice to lower the risk or rate of progression and low visual aid prescription and social support to help individuals to make the most of the vision they retain.[12] Most risk factors for AMD are non-modifiable but changes to diet, smoking cessation and exercise may lower the risk and rate of disease progression (Figure 4).

- Older age
- Presence of AMD in the other eye
- Family history of AMD
- **Smoking**
- **Hypertension**
- **Body mass index of 30Kg/m^2 or higher**
- **Diet low in omega 3 and 6, vitamins, carotenoid and minerals**
- **Diet high in fat**
- **Lack of exercise**

Figure 4. Risk factors for Age-related Macular Degeneration (AMD), with modifiable factors in bold type.[13]

The patchy loss, or atrophy, of retinal tissue that can occur in late AMD causes a corresponding loss of central visual field which can also contain the perception of scintillating light, particularly notable in dim ambient light conditions (Figure 5).

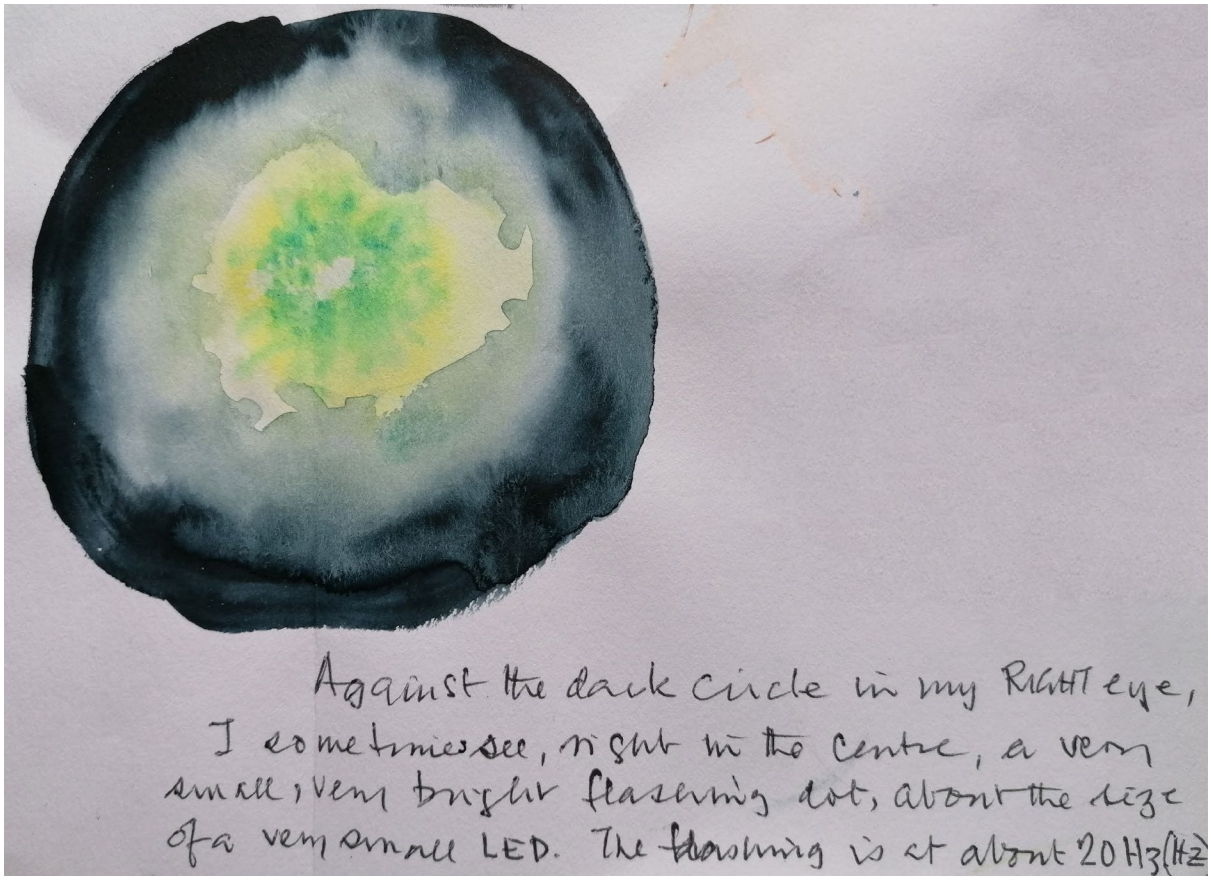


Figure 5. Annotated water colour painting volunteered by interview participant from chapter 4 to depict their central visual field and experience of scintillation. LED = Light Emitting Diode

1.1.1.1 AMD complicated by macular neovascularisation

A notable exception to the absence of disease modifying treatment for AMD is neovascular AMD (nAMD). The rapidly blinding prognosis of nAMD was transformed in 2008 by the introduction of regular injections into the eye (intravitreal injections or IVI) of anti-vascular endothelial growth factor (anti-VEGF). nAMD is thought to affect 1.4% of Europeans aged over 60.[9] It involves the pathological growth of blood vessels from the blood-rich choroid, a tissue which lies between the retina and the externally visible white sclera. The choroid usually provides oxygen and nutrients to the outer most third of the retina by diffusion alone, whilst the inner two thirds of the retina receive blood from the retinal artery, the branches of which are macroscopically visible lying 'on top' of the retina (Figure 2). The pathological choroidal vessels which characterise nAMD, grow through the normal 'blood-retina-barrier' and underneath or into the retina itself, where they leak fluid and/or blood. This can cause rapid sight loss in contrast with other forms of AMD which typically progress in a timescale of months or years. The swelling of the retina associated with this fluid leakage also leads patients' vision to become distorted and blurred, such that lines that ought to be straight are not perceived as such (Figure 6).

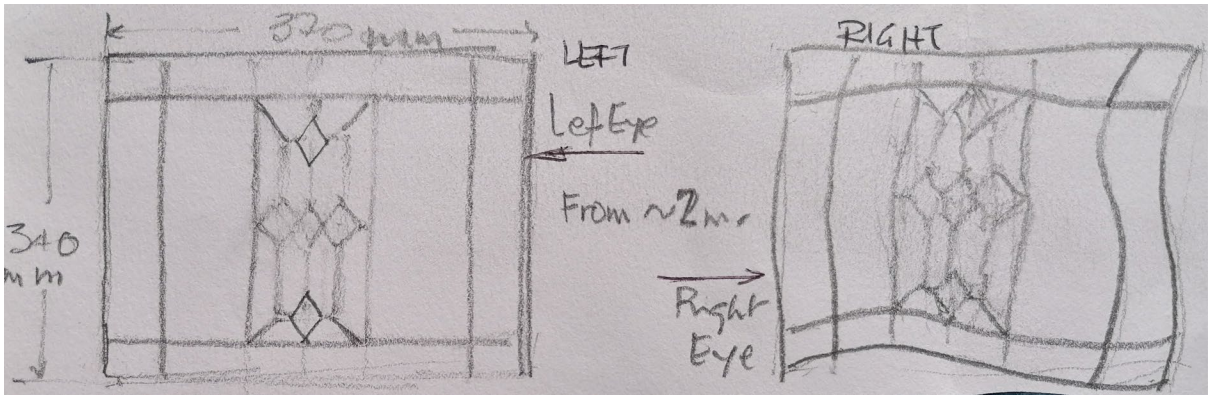


Figure 6. Annotated pencil sketch volunteered by interview participant from chapter 4 demonstrating visual distortion from their right eye, contrasting with normal perception of a stain glass window from their left eye.

1.1.1.2 nAMD treatment

Early treatment paradigms in the 1980s involved focal laser photocoagulation of the retina. This treatment damages the retina, but when this laser damage is not in the very centre of the retina, in an area known as the fovea, its visual impact is relatively mild. As a result, laser treatment for nAMD was only recommended in the minority of cases where neovascularisation occurs away from the fovea.[14] In 2004 a prospective clinical trial demonstrated that a course of anti-VEGF IVIs reduced clinical signs of nAMD and also improved vision without incurring the retinal trauma associated with laser treatment.[15] In the UK, two anti-VEGF agents were approved for clinical use in the NHS in 2008, ranibizumab and pegaptanib. This was followed by a third approved in 2013, aflibercept, and a fourth in 2021, brolicizumab.[16-18] Controversially, an unpatented and less costly first-generation anti-VEGF agent, bevacizumab, remains unlicenced for the treatment of nAMD by the Medicines and Healthcare products Regulatory Agency (MHRA). Consequently, a minority of NHS centres use it routinely despite a successful legal challenge.[19] There are hopes that some of the concerns around excessive influence and benefit from pharmaceutical stakeholders will be resolved by the on-going introduction of generic bio-similars to the second generation anti-VEGF ranibizumab following MHRA approval in 2022.[20, 21]

The major risk of intravitreal treatment (IVT) is the rare but blinding complication of infection inside the treated eye, endophthalmitis, which occurs around 1 in 2000 procedures.[22] The other disincentives to overtreatment are provider resource costs and patient inconvenience. Courses of Anti-VEGF IVIs have remained the standard of care for nAMD since their introduction in 2008. Treatment protocols to guide their use, linking the frequency of IVT to vision and imaging signs, are well established (Figure 7).[23] The continuing nature of these treatment protocols, improvements in nAMD diagnosis and an aging population have all contributed to a more than 10-fold rise in the number of IVIs over the last decade. A further doubling of nAMD IVT requirement is forecast for the decade to come and new long-term IVT regimens to treat the other form of late AMD, geographic atrophy, are expected to be approved by the National Institute for Health and Care Excellence (NICE) in 2024.[24, 25] This poses significant challenges to the limited human and financial resources available in the NHS which already struggle to meet current treatment needs.[26, 27] Real-world data from an NHS centre showed that even before the impact of the Covid-19 pandemic, delay in delivering planned treatments for nAMD was common and associated with sight loss for the patients concerned.[28] This same NHS centre had recently

displayed nAMD visual outcomes better than the mean of 12 large UK centres.[29] Given that large-scale real world evidence links a failure to deliver timely treatment to poorer visual outcomes, this suggests that such delays are widespread across the UK.[30] Since these observations, the Covid-19 pandemic has introduced additional nAMD treatment delays and avoidable sight loss, further developing the urgent need to augment clinical capacity for nAMD treatment.[31]

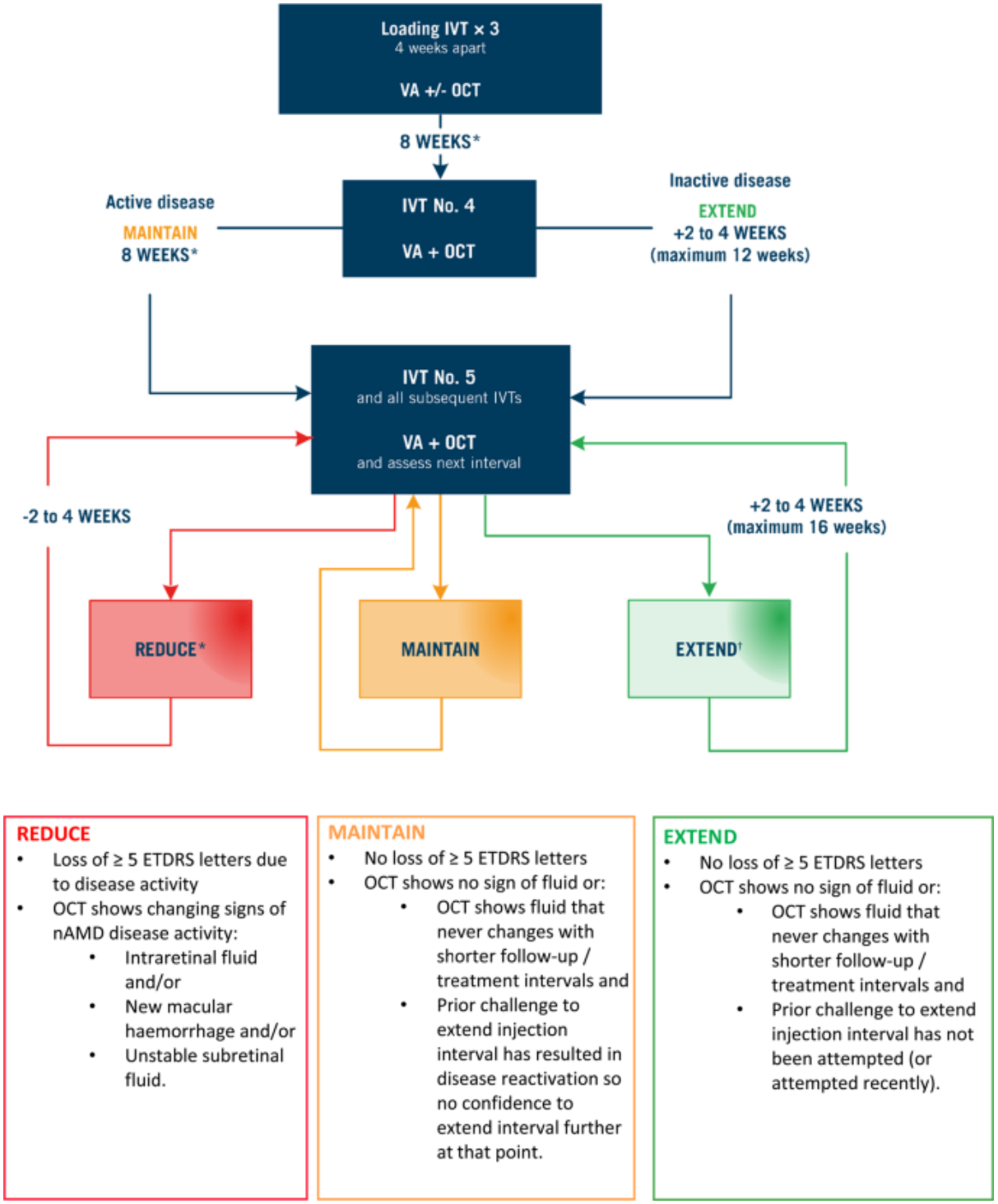


Figure 7. Figure reproduced from Ross et al, 'The recommended aflibercept T&E pathway for the treatment of patients with nAMD. IVT = Intravitreal treatment, VA = Visual Acuity, OCT = Optical Coherence Tomography, T&E = Treat and Extend, ETDRS = Early Treatment of Diabetic Retinopathy Study, nAMD = neovascular Age Related Macular Degeneration.[23]

1.1.2 Optical Coherence tomography in AMD diagnosis

Optical coherence tomography (OCT) is a non-contact imaging technique based on analysing interference patterns in emitted and reflected coherent infra-red light.[32] It is well tolerated by patients as positioning requirements are minimal, the examination takes seconds and no eye drops are required. For clinicians OCT provides micron-resolution cross-sectional imaging of the retina which can exceed the resolution required to diagnose and monitor eye diseases, even in the context of some ocular media opacity such as cataract or corneal disease. Another valuable feature is the incorporation of software which locks on to landmarks in an individual's unique retinal vasculature, ensuring that the location of imaging is reproducible over time.[33] Drusen are the key feature on OCT imaging which support the diagnosis of AMD and intraretinal fluid (IRF), subretinal fluid (SRF) and subretinal hyper-reflective material (SHRM) can all indicate the additional presence of choroidal neovascularisation, indicative of nAMD. These same features can be used to track treatment response and disease activity in nAMD over time (Figure 7). More ambiguous or complex diagnoses can be supported by other imaging modalities including OCT angiography, which can take around 10s rather than the 2s required by standard OCT. Fundus fluorescein angiography is another relevant imaging modality which requires intravenous injection of a fluorescent dye during sequential imaging.[34]

1.1.2.1 OCT in the care pathway

OCT has become a core diagnostic test in ophthalmology hospital services, particularly in retina, glaucoma and neuro-ophthalmology services.[35] Imaging equipment has become prevalent in community optometrists in recent years too, though there is no clear NHS healthcare pathway in which it sits. At present members of the public can pay a fee (usually less than £40) to have OCT imaging added into their community optometry appointment. Not all optometrists are experienced with OCT interpretation and incidental findings can lead to unnecessary referrals to hospital ophthalmology services.[36] Trials expected to report in 2024 are exploring a more established role for community-based OCT in the future, aiming to reduce the demands on hospital services and improve convenience for patients.[36-38] There are several efforts to miniaturise and cost-cut OCT imaging equipment so that private ownership and home OCT at scale become tenable.[39-41]

1.1.3 Treatment pathways for nAMD

Patient journeys with nAMD in the UK are variable but can be split into four main components (Figure 8).

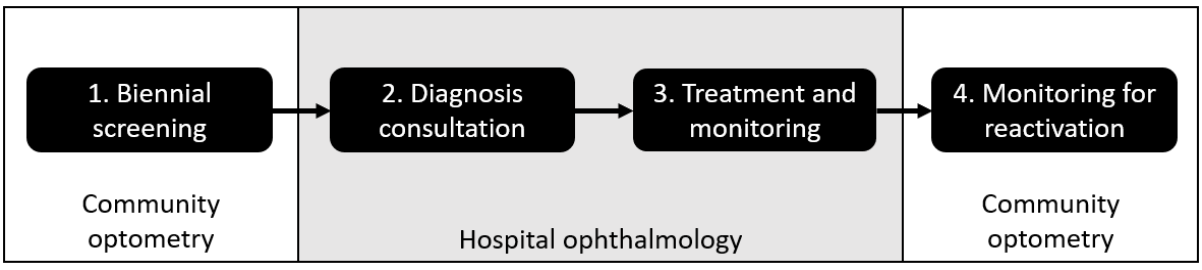


Figure 8. Stages of the patient journey for neovascular Age-related Macular Degeneration (nAMD) and their typical location.

1.1.3.1 Biennial screening in optometry services

First presentations of nAMD are common through contact with a primary care optometrist. This is either through routine eye examinations, which should take place at least biennially for all adults, or an additional patient-initiated appointment.[42] Depending on the local context, these optometrists can then request for the patient's registered general practitioner to make an urgent referral to the hospital ophthalmology service, typically the medical retina or emergency eye department services. The AMD NICE clinical guideline, updated in 2018, requires that these referrals are made within one day of presentation.[13]

Patients can also self-refer with nAMD-related visual symptoms to hospital ophthalmology services through eye emergency departments. This can also be facilitated through public facing NHS triage services, such as the 111 service, which patients may contact in the first instance. As with general emergency departments in the UK, NHS providers are incentivised to ensure patients do not wait longer than four hours to be seen.[43] Whilst it is not cost effective for hospital eye services, emergency eye department presentation often represents the quickest route to hospital ophthalmology services for members of the public.

1.1.3.2 Diagnosis and initial planning in ophthalmology services

The AMD NICE guideline requires ophthalmology services to make a diagnosis of nAMD and offer anti-VEGF treatment within 14 days of receiving a referral.[13] In addition to referrals from the community, there is also a referral contribution from other ophthalmology sub-specialty services who incidentally diagnose nAMD during the course of managing other eye conditions. The relatively high incidence of nAMD and short time scales demanded by the pathology and guideline require that hospital ophthalmology services commit significant clinical resources toward maintaining capacity to diagnose nAMD. This typically requires a consultation with the patient, retinal OCT, visual acuity (VA) assessment and slit lamp examination. When the diagnosis of nAMD is confirmed, the consultation needs to contain an explanation of the diagnosis, prognosis, potential lifestyle modification opportunities to lower the risk of fellow eye involvement and an informed consent regarding anti-VEGF treatment (Figure 7).

1.1.3.3 nAMD treatment delivery and monitoring

Depending on the length of diagnosis, whether the diagnosis is unilateral or bilateral and activity of nAMD in an eye, the interval between patients' treatment and monitoring appointments can vary from 4 – 16 weeks.[23] Pathways, staffing and the settings for the delivery of these appointments vary considerably between providers with the shortest involving IVT of anti-VEGF only and the longest also incorporating VA assessment, OCT imaging, consultation and clinical examination. NHS providers offer these services through consultant ophthalmologist led clinics, with appointments delivered by the consultants themselves, ophthalmologists at various stages of training and suitably qualified and trained allied health professionals, most commonly nurse specialists, optometrists and orthoptists.[44] All aspects are most commonly delivered face-to-face (F2F) by clinicians in a hospital setting, but increasingly pathways making use of telemedical approaches, outreach clinics and community-based IVT delivery are reported.[45] Apart from managing the interval between IVIs, this disease monitoring stage of the patient journey must also continually check the indication for treatment is still valid, screen for ocular co-morbidities and continually consider social care support interventions. These include identifying VA

below legal driving standards and initiating registration for partial or severe sight impairment.

1.1.3.4 Treatment cessation and outpatient discharge

There are no firm criteria for treatment cessation, but the joint decision is made between the clinician and patient when the remaining vision available for preservation seems disproportionate to the clinical risk, cost and inconvenience of on-going treatment. This typically occurs after AMD has caused substantial anatomical damage to the macula or when no disease activity has been noted from nAMD after periods with no IVIs, or low frequency IVIs.[13, 23] In the case where severe anatomical damage precludes the value of future IVTs, patients may be discharged from ophthalmology services immediately assuming the patient understands the rationale and the need for self-monitoring or community monitoring of the fellow eye. Where disease activity appears to have ended, an unspecified period of decreasingly frequent clinic visits without treatment are typically observed prior to discharge to the community for self-monitoring or community optometry monitoring for disease reactivation.[37]

1.1.3.5 Targeting the nAMD pathway to address the imbalance of capacity and demand

To identify where the clinical pathway for nAMD could most impactfully be targeted to redress the growing imbalance between demand and clinical capacity in the NHS, the scale, feasibility and alternatives of any innovation should be considered. Given its cyclical nature and high requirements on hospital ophthalmology service capacity, the treatment and monitoring stage of the patient journey (Figure 8) clearly represents the greatest scale of potential resource saving for the NHS. It is also true that within this stage, where sensitive discussions around diagnosis, prognosis and consent for treatment have already taken place, the decision-making process that is the main demand for clinicians' time is based on simple objective rules (Figure 7). Consequently, the treatment and monitoring stage of the patient journey seems a feasible target for some degree of automation through OCT-based clinical decision support (CDS). Such a strategy for addressing demand-capacity imbalance in retina care more broadly has also been proposed through an international Delphi process.[46] Artificial intelligence (AI)-enabled examples of such CDS have already been designed and well validated in pre-clinical settings.[47] Considering alternative clinical pathway targets, AI is already being evaluated as a non-interventional arm in prospective clinical trials to reduce the burden on hospital ophthalmology services in both the diagnosis and treatment cessation steps.[36, 37] To our knowledge, no research is published evaluating AI-enabled macula services. Therapeutics innovations aiming to extend the efficacy of each IVT are also in various stages of development and clinical evaluation but would add rather than detract value to CDS innovations.[17, 48] This is because, whilst the average treatment interval demanded by nAMD pathophysiology appears to increase there is still marked heterogeneity in treatment response between eyes. This enhances the value proposition for any low resource means of individualising treatment regimens and maintain individual safety across a population.

1.2 Artificial Intelligence

AI has been established for more than 70 years, over which time the language, science and interest has evolved.[49] At present there are many overlapping but varied definitions for AI in use by national and international authorities. According to the Organization for Economic Cooperation and Development (OECD):[50]

*'An AI system is a **machine-based system** that is capable of influencing the environment by **making recommendations, predictions or decisions** for a given set of objectives. It uses **machine and/or human-based inputs/data** to: i) perceive environments; ii) abstract these perceptions into models; and iii) interpret the models to formulate options for outcomes. AI systems are designed to operate with **varying levels of autonomy**.'*

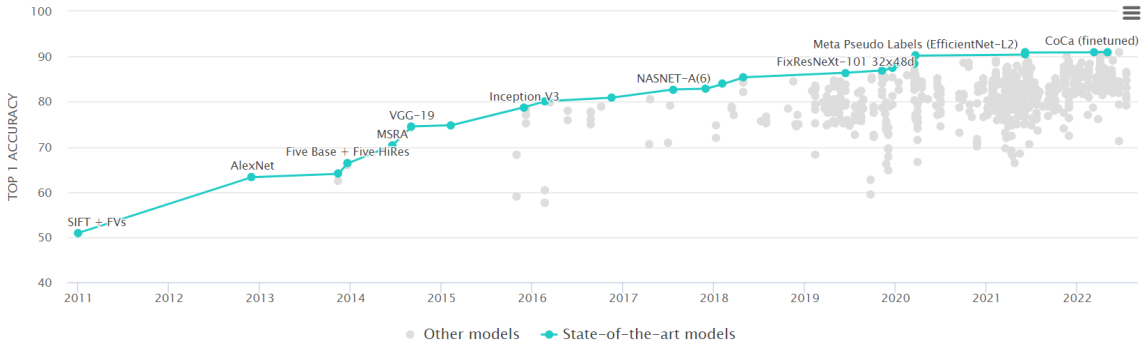


Figure 9. Screengrab from paperswithcode.com showing the percentage accuracy of state-of-the-art image classification models over time in the ImageNet Large Scale Visual Recognition Challenge.

Advances made in this field in the last decade have received a great deal of attention across society which was instigated by advances in a subset of AI techniques known as neural networks. This was exemplified by the leap in performance (Figure 9) demonstrated by a neural network known as AlexNet in a high-profile image classification competition, the ImageNet Large Scale Visual Recognition Challenge.[51]

1.2.1 Subtypes of artificial intelligence

There are several ways in which AI techniques, as set out by the OECD definition above, can be categorised. Neural networks represent a subset of AI techniques known as deep learning (Figure 10).

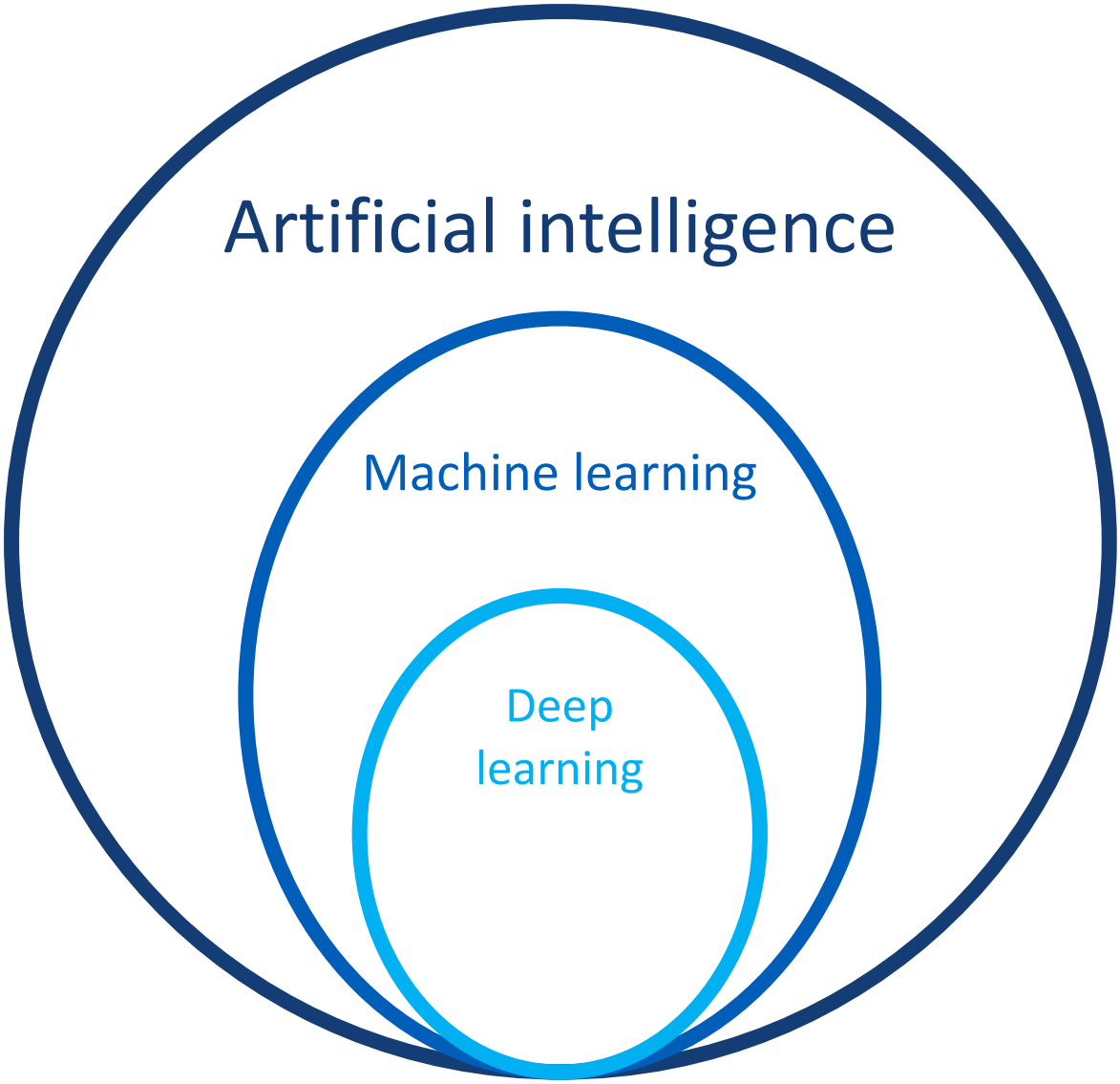


Figure 10. Venn diagram to illustrate the overlapping scope of the terms artificial intelligence, machine learning and deep learning.

1.2.1.1 Deep learning

The word 'neural' is applied to make an analogy between the mechanisms common to the computational architecture applied to deep learning technologies and the anatomy and physiology of neurones in biology. In biology each neuron is a cell which can precipitate action by maintaining a certain electrical state of polarisation or generating a wave of depolarisation which travels down cellular extensions, called axons, to another neurone or an effector tissue, like muscle.[52] This wave of depolarisation is generated when the sum of various electrical stimuli, commonly from neighbouring neurones, cross a certain threshold. In deep learning, neurones are similarly connected, with the precise number of connections, known as parameters, and layout of these connections, known as architecture, at the discretion of the computer scientist building the model (Figure 11).[53]

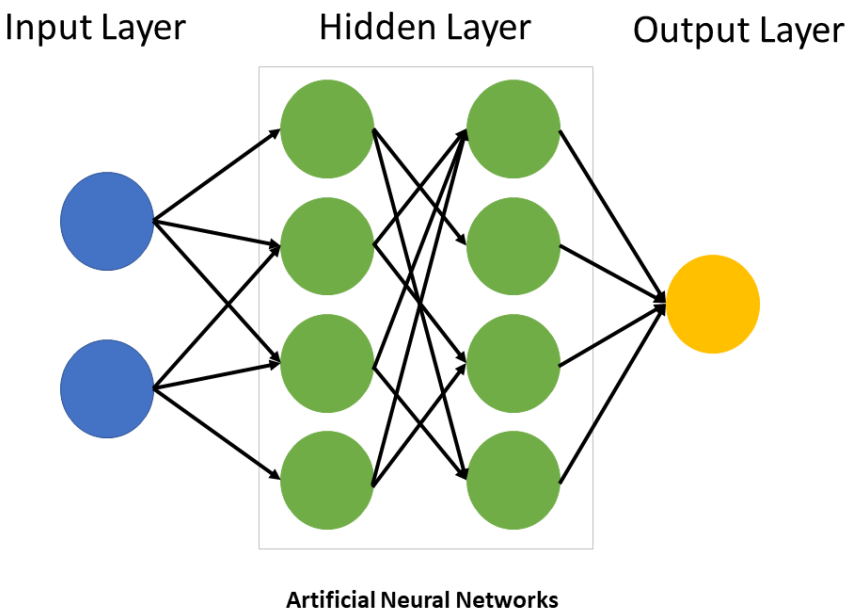


Figure 11. Schematic of neural networks showing the flow of numerical values from and input layer, through hidden layers and to an output layer.

Each neuron receives inputs from several different connections and produces a numerical value which is passed in a certain direction down its connections to neighbouring neurons. As the numerical value is passed along each of these connections that numerical value is changed by a mathematical function, known as a weight, which is specific to that connection.[53] Along with all the other numbers received as inputs to the neighbouring neuron, this numerical value is combined with those from other incoming connections to produce a new numerical value. This process of transforming and passing on numerical values continues until all the connections of the network converge on a single neuron which holds the output of the network. The learning of the network then takes place as that output is compared to a desired output through various different mechanisms. The observed disparity between the observed and desired output, known as the loss function, is used to repeatedly alter the weights of the parameters across the network with the goal of iteratively reducing the observed loss function. This mathematical process of using the loss function to iteratively improve the performance of the neural network is known as backpropagation.[53] The word 'deep' in deep learning indicates that the number of hidden layers of neurons between the input and output layers of the network is greater than one (Figure 11).

1.2.1.2 Machine learning

Similarly to the definition of AI overall, the precise demarcation of the subcategory of machine learning (ML) is disputed. ML contains all the techniques referred to as deep learning along with additional, generally less computationally demanding techniques. The mechanisms of these additional techniques also tend to be more explainable than deep learning techniques, i.e. the output can be more readily explained in terms of the way in which the input was analysed. These attributes of lower computational resource demands and greater explainability of outputs are appealing for high-risk complex applications like those in healthcare, but in certain tasks it can result in compromises in performance.[54] An

intuitive distinction can be drawn between AI techniques that do and do not satisfy the criteria for ML (and/or deep learning) by considering whether the mechanism by which inputs are processed is based on a-priori rules, derived from human knowledge, or the data itself. These tools could be distinguished as techniques that are rules or knowledge-based (AI which is not ML) or techniques which are not rules or knowledge-based (AI which is also ML).[55] These data-led ML technologies can be further categorised by the nature of the outputs they produce (i.e. categorical classifiers or continuous regressors) and the inputs which they analyse (imaging, textual, audio, tabular or multi-modal).

1.2.2 Clinical AI

AI has been applied to healthcare for several decades and has similarly been subjected to varying definitions and terms through this time.[56] These terms include Medical Expert Systems, Best Practice Alerts, Health Information Systems, CDS, Digital Health Tools, Software as a Medical Device (SaMD) and Clinical AI.[57] Applications have been made for varied purposes which can mainly be categorised as CDS, with a minority of applications forming clinical treatments in their own right, e.g. therapeutic chatbots, and more autonomous CDS which could be perceived as decision automation rather than support. This distinction between decision support and automation is both subjective and evocative. As CDS in varying forms has been prevalent in clinical practice for decades, most of the tools responsible for the on-going surge in clinical AI interest and investment could be more specifically described as deep learning enabled CDS. This term acknowledges that some elements of a deep learning enabled-CDS may depend on rule-based forms of AI, e.g. the decision to recommend antibiotics or not based upon a neural network's output regarding the probability of pneumonia from a chest x-ray.

1.2.2.1 Use cases in Ophthalmology

In part due to the prominence of diagnostic imaging in the diagnosis and management of ophthalmic disease, it is one of the most prominent medical specialties in AI applications to diagnostic imaging.[58] When considering the most recent 12 months of data on the 155 AI medical devices approved by the United States Food and Drug Administration (FDA) in from August 1st 2022 to July 30th 2023, it is clear that radiology and cardiology products dominate real-world clinical AI use, though ophthalmology is ranked fourth (Table 2).[59]

Table 2. Artificial intelligence-enabled medical devices granted approval by the United States Food and Drug Administration between August 1st 2022 and July 30th 2023, by clinical specialty for intended use.

Specialty	**n (%)**
Radiology	126 (81.3%)
Cardiovascular	12 (7.7%)
Gastroenterology	4 (2.6%)
Ophthalmology	2 (1.3%)
Neurology	6 (3.9%)
Anaesthetics	2 (1.3%)

Urology	1 (0.6%)

The absolute rate of academic publications on AI applications to ophthalmology is also rising exponentially, alongside a less smooth but equally striking increase in the rate of publication relative to all academic publishing.[60, 61]

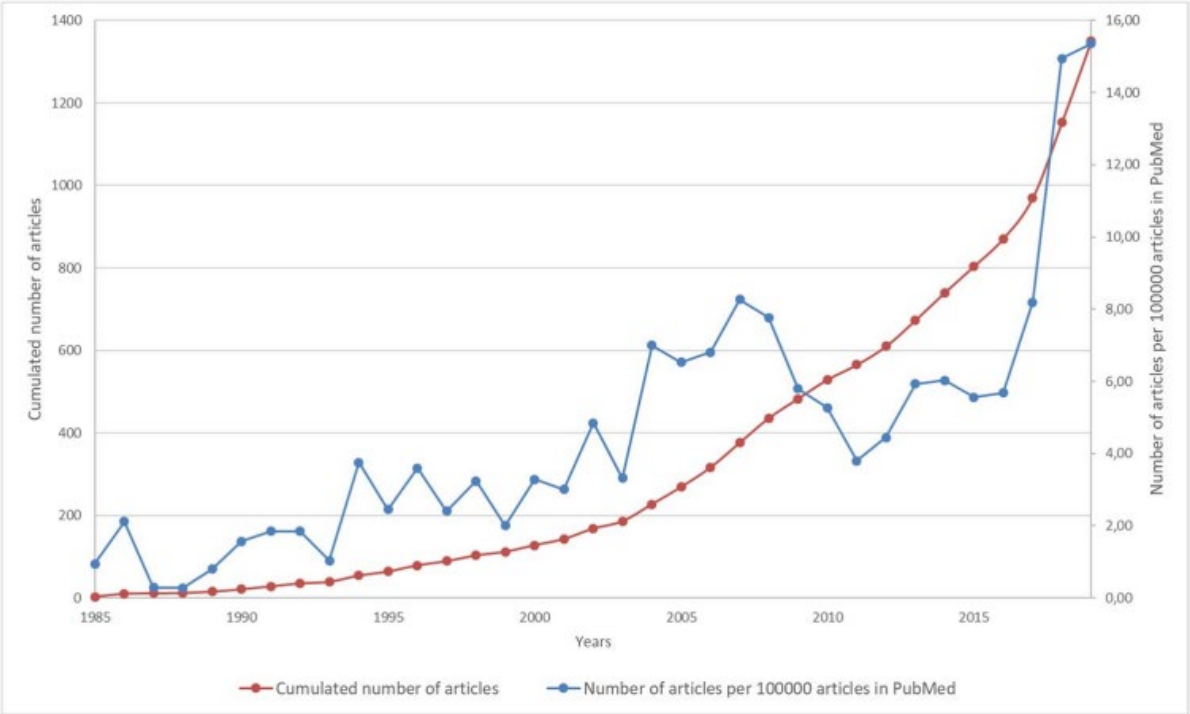


Figure 12. Growth of literature related to artificial intelligence in ophthalmology (annual number per 100,000 articles in PubMed and total number) – replicated from [60]

Disease applications within ophthalmology vary but focus on retinal disease and glaucoma, the management of which depend heavily on OCT and enface imaging modalities in standard clinical practice. Within retinal disease, AMD, other diseases affecting the macula, diabetic retinopathy and retinopathy of prematurity are the most common.[62] Strikingly, these applications tend to focus on diagnostic tasks rather than the monitoring and treatment of established chronic diseases such as glaucoma and AMD which constitute the major clinical burden on ophthalmology services (Figure 8).[5] Screening for diabetic retinopathy in enface retinal imaging is the most developed application, with fully scaled deep learning-enabled screening pathways in clinical use internationally.[63-65] Other clinical applications in diseases affecting the anterior parts of the eye include cataract, iris tumours, infectious keratitis, angle closure and keratoconus.[66]

1.2.2.2 The AI Chasm

Following the step-change in performance of AI demonstrated by AlexNet in 2012 (Figure 9), a renewed surge in policy, industry and academic interest began around deep learning-enabled CDS, which has not yet subsided.[51, 61, 67, 68] Despite this broad optimism, there are only a handful of examples of application in real-world NHS care, which reflects a know-do gap referred to as the 'AI chasm'.[69, 70] As such, despite promising pre-clinical performance, many implementations have proved operationally infeasible whilst others have found the new technology to detract from the current standard of care.[71]

To resolve this disconnect between the expectations assigned to deep learning-enabled CDS and the reality for real-world care, key stakeholders must share their own perspectives and understand others' on the factors which sustain this AI-chasm. Many factors have been well characterised in prior syntheses regarding CDS use, but there appear to be novel factors which specifically influence the implementation of deep learning enabled CDS.[56, 57] These factors are in part inherent to the technology itself, but also arise from how potential adopters perceive it and the wider organisational, social and cultural contexts in which it must be implemented.[72] Examples include:

- the regulatory landscape in which AI implementation must take place is evolving fast relative to the years it takes to develop a potential AI use case into a product ready to me implemented. This is exemplified by the recently published AI Act from the European Parliament.[73]
- Workforce readiness is another factor which challenges the implementation of AI. This has been identified in NHS policy documents for at least 5 years, but still remains an active concern with few clear solutions.[74,75]
- The evidence needs for adoption are largely unclear or unmet. Chief among these is economic evidence, which is critical for Health Technology Assessment agencies like NICE, but also decision makers within potential adopting organisations reviewing business cases. The economic evidence available so far across clinical AI appears patchy with differing conclusions across use cases.[76,77]

The literature describing these distinguishing factors for deep learning enabled CDS is relatively scarce and often contains little scientific rigour. However, informative empirical investigations are beginning to be published from a handful of deep learning enabled CDS use cases, providing valuable insights for proponents of their successful implementation.[64, 78] This literature is developing slowly and represents heterogeneous tools, adopters and contexts across distinct implementation use cases. It will therefore be important to understand if, when and how learnings for specific use cases can be leveraged to support the implementation of different deep learning enabled CDS to different healthcare niches.

1.3 Implementation science

Implementation science is a relatively young disciplinary field which focuses on how and why interventions work, rather than testing whether they work or not.[79] It achieves this through the application of theoretically-informed and empirical research (Figure 13).[80] Whilst the schematic below suggests that implementation research follows on sequentially from effectiveness research, it is often hybridised with effectiveness research to increase the efficiency through which research can translate interventions into practice.[81] At one extreme this can take the form of research primarily focused on evidencing the effectiveness of an intervention, but accompanied by non-interventional data collection and analysis to understand factors that could influence implementation. At the other, an intervention with clear evidence of effectiveness can undergo a clinical trial where the difference between intervention and comparator arms is simply the implementation strategies which are deployed around the intervention.

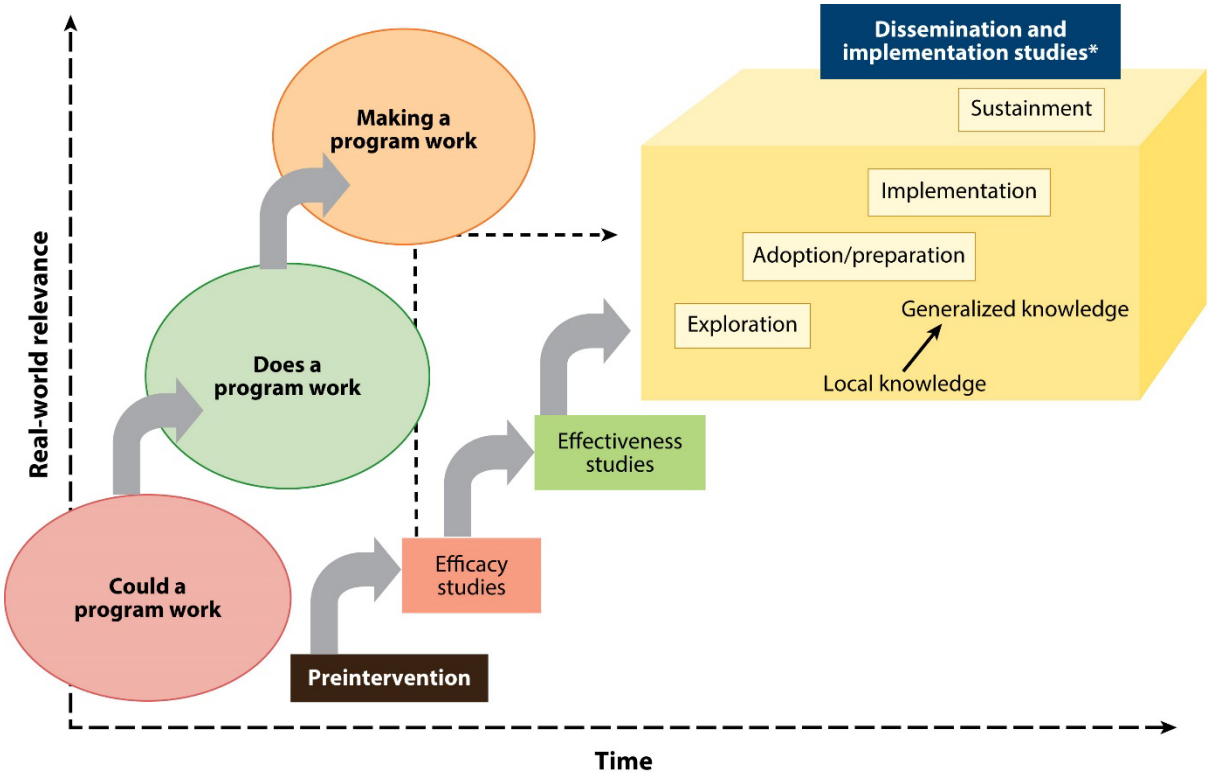


Figure 13. Schematic of research pipeline and the positioning of implementation research within it reproduced from [80]

To close 'know-do gaps' in healthcare, such as the 'AI-chasm' a multi-disciplinary approach and across a spectrum of research and practice is required.[70, 82] To ensure scalable value, the research involved must also create insights that are relevant across varied innovations and contexts, which requires a degree of abstraction. Increasing abstraction helps to make insights from one setting at least partly generalisable to others. Decreasing abstraction, helps to transform these high-level insights into practical and actionable knowledge for a specific situation. This process of abstraction, analogous to the underpinnings of abstract art, necessitates an ability to remove or add detail whilst minimising the loss of meaning (Figure 14). Ensuring the validity of these abstractions, which can be thought of as the process of theorizing, has attracted input from various academic disciplines and their accompanying breadth of research paradigms. Attempts to reconcile or promote differing perspectives from these paradigms has been a persistent source of agitation within the field of implementation science. This agitation may progress the science itself, but often threatens the desired accessibility of the field.[83, 84] This is well demonstrated by the 113 distinct theories, models or frameworks (TMFs) catalogued in a single peer-reviewed library, which is merely a fraction of published TMFs relevant to implementation science which have been prioritised for peer-review and curation.[85]

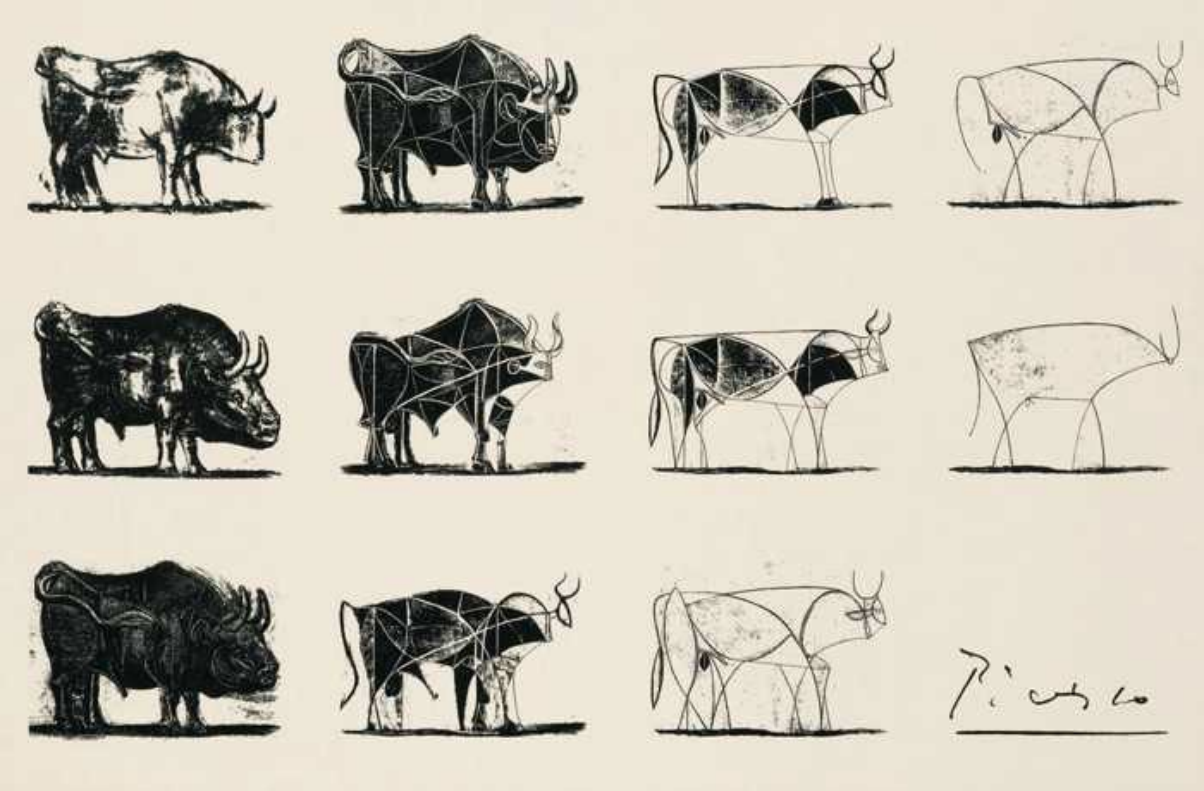


Figure 14. 'Bull', 1945; a series of 11 lithographs by Pablo Picasso. As in the process of theorising, the essence of a phenomenon (in this case the bull) is accurately represented with varying degrees of abstraction.[86]

1.3.1 Theory and implementation science

Unlike its interface with adjacent and overlapping fields, the importance of TMFs within implementation science is uncontested and a good starting point in characterising the field. Despite that, the meaning attributed to language applied to TMFs is not standardised, but throughout this thesis the acronym TMF will be used as an umbrella term to refer to theories and/or frameworks and/or models. TMFs are directly lifted or adapted from diverse fields of research including psychology, public health, social science, healthcare, business, organisational theory and political science.[85] They are also developed anew from empirical observations of implementation. Given the open forum for dialogue between disciplines and the realms of theory and practice which implementation seeks to provide, classifying TMFs by their academic origins may be possible but seems counterproductive.[87] Such disciplinary distinctions also do little to inform any given TMF's application. Per Nilsen's taxonomy of TMFs used in implementation science (Figure 15) represents a more useful and emollient classification system as a way to guide the application of particular TMFs in a given context.[88] Other than this prominent example of categorisation by purpose, another helpful system by which to categorise TMFs by is the scope of their relevance.

1.3.1.1 TMFs by purpose

Nilsen categorises theoretical application in implementation science into 3 purposes (Figure 15):

1 To illuminate the process and mechanisms by which research is translated into practice – the 'physiology' of implementation

2 To understand the factors and interdependencies which may influence implementation – the 'anatomy' of implementation
3 To structure the evaluation of implementation endeavours that are past or planned – an implementation scorecard

TMFs in the first category are termed process models and are particularly useful in understanding why aspects of an implementation exercise may or did produce certain results, e.g. the Technology Acceptance Model (TAM).[89] A particularly practical sub-type of the process model aims to prescribe actions for the actors in a certain implementation exercise, and are known as action models.

The second category is subdivided into three; determinant frameworks (2a), classic theories (2b) and implementation theories (2c). Determinant frameworks seek to identify the factors that may influence the successful implementation of an intervention. This supports comparison across different implementation settings without risking false assumptions about the importance or mechanisms of influence of a given factor. Classic theories are from fields which would broadly be considered as distinct from implementation science but are judged to hold value for the process of implementation., e.g. Rogers' Theory of diffusion.[90] Implementation theories are TMFs which have been consciously adapted or developed for application within the field of implementation science, e.g. the Consolidated Framework for Implementation Research.[91] Deciding whether a TMF was specifically developed for application within the disciplinary field of implementation science is no less contentious than the definition of implementation science itself. Certain TMF contributors would nevertheless identify with this approach to categorisation.

The final category of TMF purposes is that of evaluation frameworks. The TMFs here are likely to co-exist in one of the other categories but are defined by their use to assess planned or executed implementation strategies.

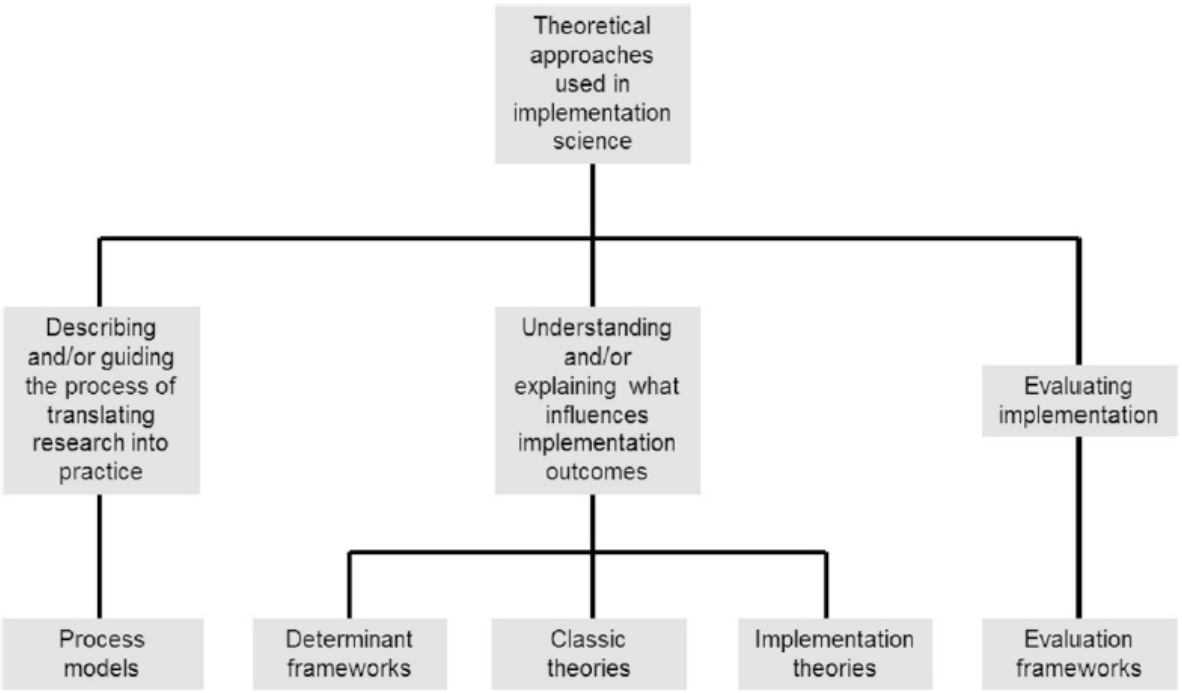


Figure 15. A Direct reproduction of Per Nilsen's 2015 taxonomy of theoretical approaches used in implementation science, divided into three distinct purposes of a theoretical approach and five named categories.[88]

There are some criticisms to be made of Nilsen's taxonomy, most notably that there is little consensus over which disciplines lie within and without implementation science yet the distinction between categories 2b and 2c depends upon it.[92] Any given TMF could also be attributed to several of the categories depending on how it is being applied. However, this context specific classification could be considered a strength rather than a limitation as it better reflects the versatility of TMFs within implementation science. More prescriptive and granular categorisations of the purposes to which TMFs are applied in implementation science have also been published.[93]

1.3.1.2 TMFs by scope of relevance

TMFs represent an abstraction of empirical observations which permit greater generalisability and accessibility of the value within those observations (Figure 14). TMFs could be defined as "an ordered set of assertions about a generic behaviour or structure assumed to hold throughout a significantly broad range of specific instances".[94] In considering this definition, the question of 'how broad is significantly broad?' seems unavoidable and the answer must vary with the desired application of a TMF (Figure 16). On one extreme, high degrees of abstraction can be applied to observations to produce assertions that appear valid over a large range of instances but are incapable of illuminating factors and interactions which are highly specific to a given context. At the other extreme, lower degrees of abstraction can be applied to observations, which compromises the generalisability of a TMF but increases the detail of the insights which may be offered to the instances which lie within its scope. This spectrum of abstraction has been arbitrarily sliced into three categories of programme, mid-range and grand TMFs (referred to as programme theory, mid-range theory and grand theory by authors).[95] The complementary values of taking broad insights from diverse contexts and more specific insights from contexts similar to the instance of interest, is part of what has incentivised the great number of published

TMFs. TMFs which facilitate theoretically informed practice within implementation science often fall within the loose category of 'mid-range'. This prevents the extent of TMFs catalogued in the literature from becoming too overwhelming but maintains clear relevance between published TMFs and various applications.

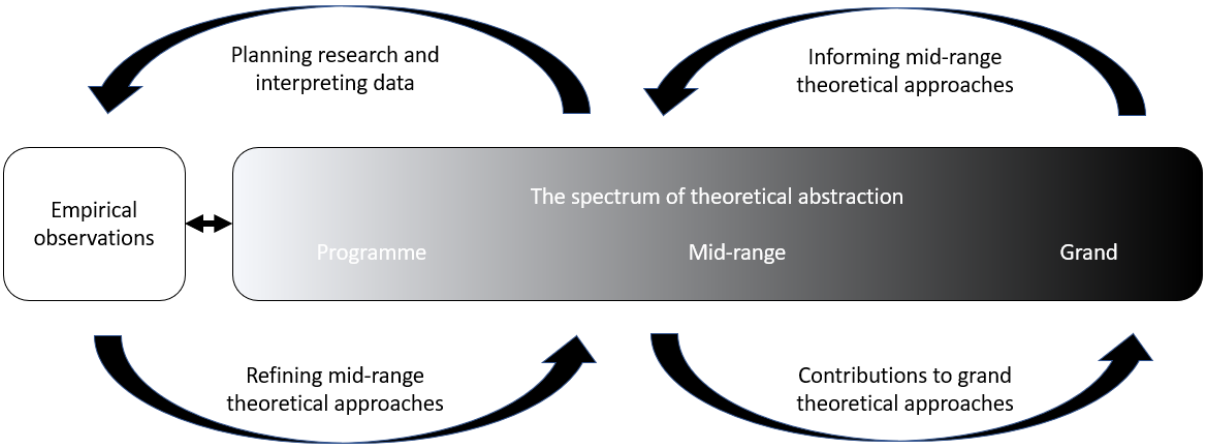


Figure 16. Theorising in implementation science, adapted from [95]

Through application a TMF will often be adapted using empirical observations from a particular context, perspectives from relevant stakeholders and other relevant pre-existent TMFs. Researchers can then reflect on the value of the adaptions they made to pre-existent mid-range TMFs within their programme theory and share these insights through publication. This process of adaption produces another characteristic of implementation science, not just to be theoretically informed, but also to be theoretically informative.[95] In reality, this iterative refinement and evolution of TMFs is often disrupted by the unavoidably limited oversight that researchers and practitioners have of all available TMFs. Alongside the genuine value of diverse TMFs for diverse implementation challenges, these shortcomings help to explain the proliferation of TMFs.

1.3.2 Implementation practice

As a field of research, implementation science is the final stage on the translation pathway, yet there is still some distinction between the individuals and activities associated with its research and practice.[96, 97] For implementation science research, there is greater focus on TMFs, theorising and an understanding of various research methods which have been described above and in the methods sections across this thesis. In implementation practice, there is a greater focus on the strategies which help to translate understanding and theory into action, the measurement of outcomes to allow an agile and responsive approach to implementation and the ability to identify and engage stakeholders in the process.[97]

1.3.2.1 Implementation strategies

Stakeholders engaged more closely with the world of clinical practice rather than academia often respond with some synonym of 'so what!?' when presented with even the most rigorous and accessible analyses of empirical qualitative data. One of the means of bridging this disconnect between research and practice is through the application of implementation strategies. Implementation strategies aim to translate knowledge about factors that influence implementation into pragmatic actions which increase practitioners' probability of achieving their context specific definitions of success. In 2015 a multi-institutional collection

of North-American implementation experts participated in a rigorous consensus exercise aiming to produce an exhaustive list of distinct strategies, relevant across innovations and contexts.[98] This produced a list of 73 strategies with accompanying definitions, which were subsequently grouped into 9 categories and rated for relative importance and feasibility in the Expert Recommendations for Implementing Change (ERIC) study.[98] Similarly to the process of abstraction described above relating to TMFs (Figure 14), these strategies are described in such a way that makes them generalisable and accessible to a wide range of contexts and stakeholders.

1	Engage consumers
2	Use evaluative and iterative strategies
3	Change infrastructure
4	Adapt and tailor to the context
5	Develop stakeholder interrelationships
6	Utilise financial strategies
7	Support clinicians
8	Provide interactive assistance
9	Train and educate stakeholders

Figure 17. Nine categories of 73 implementation strategies synthesised from the literature then consolidated and categorised through expert consensus.[98]

The intention for practitioners is not to seek to leverage all 73 strategies to improve their chances in each implementation endeavour, but rather to treat it as a library to consult with their own practical and theoretical insights to prioritise approaches they wish to take.[99]

1.3.2.2 Measurement of implementation outcomes and mechanisms

In recent decades the focus in biomedical research has shifted away from establishing the efficacy on innovations, 'does it work?', to questions surrounding effectiveness, 'does it work here and now?'.[79] Sequential iterations of the UK Medical Research Council's framework for the evaluation of complex interventions have taken this further toward measuring the mechanisms by which an innovation is or is not effective and also the influences it exerts over the host system.[79] This demands the collection of highly varied datapoints, often in parallel to an active implementation process. Ideally, these datapoints would meaningfully reflect factors that are expected to be influential or implementation strategies which have been selected. This accommodating and adaptive approach has been coined as a fourth research paradigm, with the preceding stages of quantitative, qualitative and mixed-methods research.[100] Here research takes place in real-time and in partnership between those 'conducting' and 'participating' in the research with no sense of methodological superiority or inferiority, simply the fit between method, context and actors.

As with other aspects of implementation science, the prospect of such flexibility and context-focused research is exciting, but a little intimidating for any individual charged with the responsibility of making sense and value in such complexity. To complement aforementioned libraries of TMFs and implementation strategies, there are also searchable tools and scales to measure various outcomes and mechanisms that are established within implementation science.[101] Whilst these methodological shortcuts pose some risk of

misapplication and misinterpretation through inadequate training and experience of users, they have the pragmatic aim of making implementation research more accessible to communities of practitioners.

1.3.3 Applying implementation science to AI-enabled macula services

An analysis of abstracts accepted by the Royal College of Ophthalmologists for presentation at their annual congress shows just 11.3% contain qualitative methods of any form, only 1.5% of which represent interview, focus group or observation.[102] Recent qualitative evidence syntheses suggest a similarly limited qualitative representation in the clinical AI literature.[103, 104] The combination of these two limitations means that very few insights into the complexity surrounding innovations related to AI-enabled macula services are available.[105] Whilst such directly relevant evidence would be useful, its absence emphasises the value of implementation science in abstracting insights from differing contexts to pragmatically inform implementation efforts for AI-enabled macula services.

The evidence and regulatory foundations required to optimise AI-enabled macula services has not yet been established. Most notably, current research is non-interventional effectiveness research. This necessitates a more exploratory tone to implementation research, which will be hybridised with effectiveness research. Considering the ImpRes toolkit, a rigorously curated set of implementation research methods, this should include stakeholder involvement, identifying determinants of implementation, and measures for both implementation and clinical outcomes.[106] This holistic evidence generation will establish if the investment necessitated by an interventional evaluation of AI-enabled macula services is justified together with the design of the intervention which seems most likely to deliver success for the NHS and its patients.

Figure 18. Schematic of the ImpRes toolkit reproduced from [106]

1.4 Statement of the problem

This thesis aims to explore the factors which sustain the AI chasm in healthcare generally and in AI-enabled macula services in the NHS. Recommendations to support resolution of the AI chasm for research and practice will be made from this exploration of healthcare as a whole. For AI-enabled macula services, where feasible evidence will be generated to support the resolution of the barriers to implementation that are identified. The thesis will achieve these aims by addressing five distinct evidence gaps across chapters 3 - 7:

1. A meaningful qualitative evidence synthesis of clinical AI implementation research is not available to inform implementation research and practice (chapter 3).
2. The factors that could influence the implementation of an AI technology within nAMD clinical pathways have not been explored (chapter 4).
3. The ability of AI as a medical device (AIaMD) to meet nAMD service stakeholders' minimum requirements for acceptance is untested (chapter 5).

4. A healthcare intervention to deliver the AIaMD into macula services has not been designed to align with the factors likely to influence implementation (chapter 6).
5. A systematic consideration of risks that an AI-intervention for treatment monitoring may pose across the nAMD care pathway, and their potential mitigation, has not been performed (chapter 7).

Whilst economic evidence is likely to support the implementation of AI-enabled macula services it is not considered in scope for this thesis as it depends upon an as yet undefined intervention and substantial additional time and expertise. The thesis will however propose an intervention suitable for economic evaluation and further refinement if it should appear advantageous. Even without economic evaluation, delivering a programme of work to address the evidence gaps outlined above will require the application of multiple research methods. To maximise the transparency of its findings and the coherence of the methods chosen, chapter 2 will establish an overarching approach to guide its design and conduct.

1.5 Structure of the thesis

Beyond this introduction there are 7 further chapters. Chapter 2 discusses the over-arching philosophical approach for the thesis. Chapter 3 systematically synthesises qualitative research of clinical AI to characterise key stakeholder groups, determinants of implementation and TMF applications. These insights inform the design of a primary qualitative study in chapter 4, recruiting key stakeholders of a macula service to identify *what* could influence the implementation of AI-enabled macula services and *why*. These stakeholders prioritise vision preservation, and so chapter 5 tests the non-inferiority of AI-enabled assessments of nAMD activity against consultant-led-care (CLC) by simulating a potential medical device in which the AI technology could be embedded. The positive result of chapter 5 signals potential acceptability of such a medical device, but highlights the need for an understanding of *how* to conduct the implementation of AI-enabled macula services. Chapter 6 performs a theory-informed secondary analysis of data from chapter 4, to design an actionable and evidence-based AI-enabled intervention. This proposed intervention permits an evaluation of a full hypothetical AI-enabled healthcare pathway in chapter 7, using methods familiar to key decision makers in clinical AI implementation. Besides producing evidence in a meaningful format to relevant practitioners, this enables recommendations to further improve service and implementation outcomes. Chapter 8 discusses the thesis' findings in the wider research and practice context and summarises its conclusions.

1.6 Statement on ethics

A protocol for the work contained within this thesis and associated documentation was submitted for ethical review through the Integrated Research Application System portal (IRAS project ID 280448). Newcastle upon Tyne Hospitals NHS Foundation Trust was the sponsor for the research. The project received initial approval from the North West – Greater Manchester South Research Ethics Committee on 12th May 2021 (REC reference 21/NW/0138). Initial approval was received from the NHS Health Research Authority and Health and Care Research Wales on 17th May 2021. Subsequently two amendments were submitted and came to be approval by the sponsor, Research Ethics Committee and Health Research Authority. The first (non-substantial) amendment permitted separate consent forms for patient and carer participants rather than a shared form for both. The second (substantial) amendment permitted the use of a more operationally convenient approach to

data egress between Newcastle upon Tyne Hospitals and Moorfields Eye Hospital NHS Foundation Trusts.

1.7 References

1. The Royal College of Ophthalmologists, Facing workforce shortages and backlogs in the aftermath of COVID-19: The 2022 census of the ophthalmology consultant, trainee and SAS workforce. 2023.

2. The Royal College of Ophthalmologists, Launch of physician associate pilot. 2023.

3. The Royal College of Ophthalmologists, Ophthalmology Local Training (OLT) Programme. 2021.

4. The Royal College of Ophthalmologists, RCOphth completes Specialty Specific Guidance for New GMC Standard. 2023.

5. NHS Digital, Hospital Episode Statistics. 2023.

6. Foot, B. and C. MacEwen, Surveillance of sight loss due to delay in ophthalmic treatment or review: frequency, cause and outcome. Eye (Lond), 2017. 31(5): p. 771-775.

7. The Association of Optometrists, NHS patient backlogs are leading to life-changing sight loss, FOI request reveals. 2023.

8. Bourne, R., et al., Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. The Lancet Global Health, 2021. 9(2): p. e130-e143.

9. Li, J.Q., et al., Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis. Br J Ophthalmol, 2020. 104(8): p. 1077-1084.

10. The Royal National Institute for the Blind, Sight loss data tool. 2023.

11. Ferris, F.L., 3rd, et al., Clinical classification of age-related macular degeneration. Ophthalmology, 2013. 120(4): p. 844-51.

12. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E and beta carotene for age-related cataract and vision loss: AREDS report no. 9. Arch Ophthalmol, 2001. 119(10): p. 1439-52.

13. The National Institute for Health and Care Excellence, NICE guideline 82 - Age-related macular degeneration. 2018.

14. Argon laser photocoagulation for neovascular maculopathy. Three-year results from randomized clinical trials. Macular Photocoagulation Study Group. Arch Ophthalmol, 1986. 104(5): p. 694-701.

15. Gragoudas, E.S., et al., Pegaptanib for Neovascular Age-Related Macular Degeneration. New England Journal of Medicine, 2004. 351(27): p. 2805-2816.

16. The National Institute for Health and Care Excellence, Aflibercept solution for injection for treating wet age-related macular degeneration. 2013.

17. The National Institute for Health and Care Excellence, Brolucizumab for treating wet age-related macular degeneration. 2021.

18. The National Institute for Health and Care Excellence, Ranibizumab and pegaptanib for the treatment of age-related macular degeneration. 2008.

19. Cohen, D., CCGs win right to offer patients Avastin for wet AMD. BMJ, 2018. 362: p. k4035.

20. Neil, M.B., et al., Biosimilar SB11 versus reference ranibizumab in neovascular age-related macular degeneration: 1-year phase III randomised clinical trial outcomes. British Journal of Ophthalmology, 2023. 107(3): p. 384.

21. Agency, M.a.H.R., Summary of Product Characteristics - Ongavia. 2022.

22. Kiss, S., et al., Endophthalmitis rates among patients receiving intravitreal anti-VEGF injections: a USA claims analysis. Clin Ophthalmol, 2018. 12: p. 1625-1635.

23. Ross, A.H., et al., Recommendations by a UK expert panel on an aflibercept treat-and-extend pathway for the treatment of neovascular age-related macular degeneration. Eye, 2020. 34(10): p. 1825-1834.

24. Chopra, R., et al., Intravitreal injections: past trends and future projections within a UK tertiary hospital. Eye, 2022. 36(7): p. 1373-1378.

25. The United States Food and Drug Administration, FDA Approved Drugs: New Drug Application (NDA): 217171. 2023.

26. Hollingworth, W., et al., A longitudinal study to assess the frequency and cost of antivascular endothelial therapy, and inequalities in access, in England between 2005 and 2015. BMJ Open, 2017. 7(10): p. e018289.

27. Gale, R., et al., Health technology assessment of new retinal treatments; the need to capture healthcare capacity issues. Eye, 2022. 36(12): p. 2236-2238.

28. Hogg, J. The prevalence and impact of treatment delays in exudative age-related macular degeneration. 2021. Investigative Ophthalmology & Visual Science.

29. Talks, J.S., et al., Appropriateness of quality standards for meaningful intercentre comparisons of aflibercept service provision for neovascular age-related macular degeneration. Eye (Lond), 2017. 31(11): p. 1613-1620.

30. Fu, D.J., et al., Insights From Survival Analyses During 12 Years of Anti–Vascular Endothelial Growth Factor Therapy for Neovascular Age-Related Macular Degeneration. JAMA Ophthalmology, 2021. 139(1): p. 57-67.

31. Arruabarrena, C., et al., Impact on Visual Acuity in Neovascular Age Related Macular Degeneration (nAMD) in Europe Due to COVID-19 Pandemic Lockdown. J Clin Med, 2021. 10(15).

32. Huang, D., et al., Optical coherence tomography. Science, 1991. 254(5035): p. 1178-81.

33. Coscas, G., et al, Heidelberg Spectralis Optical Coherence Tomography Angiography: Technical Aspects. Dev Ophthalmol, 2016. 56: p. 1-5.

34. Niestrata, M., et al., ATHENA Study Protocol - Optical Coherence Tomography Angiography Analysis for the detection of Neovascular Age-related Macular Degeneration: a Comprehensive Multi-Centre, Prospective, Randomised Diagnostic Accuracy and Non-Inferiority Study. Investigative Ophthalmology & Visual Science, 2022. 63(7): p. 2902 – F0055-2902 – F0055.

35. Pontikos, N., et al., Ten Years of Optical Coherence Tomography in Ophthalmology: Current and Future Use. Investigative Ophthalmology & Visual Science, 2018. 59(9): p. 3216-3216.

36. Ji Eun Diana, H., et al., Teleophthalmology-enabled and artificial intelligence-ready referral pathway for community optometry referrals of retinal disease (HERMES): a Cluster Randomised Superiority Trial with a linked Diagnostic Accuracy Study—HERMES study report 1—study protocol. BMJ Open, 2022. 12(2): p. e055845.

37. Annastazia, E.L., et al., FENETRE study: quality-assured follow-up of quiescent neovascular age-related macular degeneration by non-medical practitioners: study protocol and statistical analysis plan for a randomised controlled trial. BMJ Open, 2021. 11(5): p. e049411.

38. Barnaby, C.R., et al., Effectiveness of Community versus Hospital Eye Service follow-up for patients with neovascular age-related macular degeneration with quiescent disease (ECHoES): a virtual non-inferiority trial. BMJ Open, 2016. 6(7): p. e010685.

39. Liu, Y., N.M. et al, Prospective, Longitudinal Study: Daily Self-Imaging with Home OCT for Neovascular Age-Related Macular Degeneration. Ophthalmology Retina, 2022. 6(7): p. 575-585.

40. Kim, J.E., et al., Evaluation of a self-imaging SD-OCT system designed for remote home monitoring. BMC Ophthalmology, 2022. 22(1): p. 261.

41. von der Burchard, C., et al., Self-examination low-cost full-field OCT (SELFF-OCT) for patients with various macular diseases. Graefe's Archive for Clinical and Experimental Ophthalmology, 2021. 259(6): p. 1503-1511.

42. The College of Optometrists, Guidance for Professional Practice; Knowledge Skills and Performance. 2023.

43. Parkin, E., NHS maximum waiting time standards; Briefing paper, H.o.C. Library, Editor. 2020.

44. Hasan, H., S. et al, Setting up a successful nurse-led intravitreal injections service: pearls from Swindon. Br J Nurs, 2020. 29(20): p. 1178-1185.

45. Amoaku, W., et al., Providing a Safe and Effective Intravitreal Treatment Service: Strategies for Service Delivery. Clin Ophthalmol, 2020. 14: p. 1315-1328.

46. Loewenstein, A., et al., Save our Sight (SOS): a collective call-to-action for enhanced retinal care across health systems in high income countries. Eye (Lond), 2023. 37(16): p. 3351-3359.

47. De Fauw, J., et al., Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med, 2018. 24(9): p. 1342-1350.

48. Holekamp, N.M., et al., Archway Randomized Phase 3 Trial of the Port Delivery System with Ranibizumab for Neovascular Age-Related Macular Degeneration. Ophthalmology, 2022. 129(3): p. 295-307.

49. Turing, AM, COMPUTING MACHINERY AND INTELLIGENCE. Mind, 1950. 49(236): p. 433 - 460.

50. The Organization for Economic Cooperation and Development, Recommendation of the Council on Artificial Intelligence. 2023.

51. Krizhevsky, A., et al, ImageNet classification with deep convolutional neural networks, in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. 2012, Curran Associates Inc.: Lake Tahoe, Nevada. p. 1097–1105.

52. Mills, K.R., Oxford Textbook of Neurophysiology. 2017: Oxford University Press.

53. LeCun, Y., Y. Bengio, and G. Hinton, Deep learning. Nature, 2015. 521(7553): p. 436-444.

54. Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019. 1(5): p. 206-215.

55. Sutton, R.T., et al., An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digital Medicine, 2020. 3(1): p. 17.

56. Kawamoto, K., et al., Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj, 2005. 330(7494): p. 765.

57. Miller, A., et al., Integrating computerized clinical decision support systems into clinical work: A meta-synthesis of qualitative research. Int J Med Inform, 2015. 84(12): p. 1009-18.

58. Liu, X., et al., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health, 2019. 1(6): p. e271-e297.

59. The United States Food and Drug Administration, Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. 2023.

60. Boudry, C., et al., Analysis of international publication trends in artificial intelligence in ophthalmology. Graefe's Archive for Clinical and Experimental Ophthalmology, 2022. 260(5): p. 1779-1788.

61. Zhang, J., et al., An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. The Lancet Digital Health, 2022. 4(4): p. e212-e213.

62. Ting, D.S.W., et al., Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol, 2019. 103(2): p. 167-175.

63. Peter, H., et al., Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. British Journal of Ophthalmology, 2021. 105(5): p. 723.

64. Beede, E., et al., A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020, Association for Computing Machinery: Honolulu, HI, USA. p. 1–12.

65. Savoy, M, IDx-DR for Diabetic Retinopathy Screening. American Family Physician, 2020. 101(5): p. 307-308.

66. Darren Shu Jeng, T., et al., Artificial intelligence for anterior segment diseases: Emerging applications in ophthalmology. British Journal of Ophthalmology, 2021. 105(2): p. 158.

67. Muehlematter, U.J., et al, Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health, 2021. 3(3): p. e195-e203.

68. Health Ethics and Governance team, W.H.O., Ethics and governance of artificial intelligence for health, World Health Organization, Editor. 2021.

69. Yin, J., K.Y. Ngiam, and H.H. Teo, Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. J Med Internet Res, 2021. 23(4): p. e25759.

70. Topol, E.J., High-performance medicine: the convergence of human and artificial intelligence. Nat Med, 2019. 25(1): p. 44-56.

71. Lin, H., et al., Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. eClinicalMedicine, 2019. 9: p. 52-59.

72. Maniatopoulos, G., et al., Moving beyond local practice: reconfiguring the adoption of a breast cancer diagnostic technology. Soc Sci Med, 2015. 131: p. 98-106.

73. European Parliament. REGULATION (EU) 2024/… OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of … laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797

and (EU) 2020/1828 (Artificial Intelligence Act).
https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

74. NHS England, The Topol Review. 2019.

75. NHS AI Lab, NHS Transformation Directorate and NHS Health Education England, Developing healthcare workers' confidence in artificial intelligence (AI) (Part 2). 2022.

76. Bharadwaj P, et al. Unlocking the Value: Quantifying the Return on Investment of Hospital Artificial Intelligence. J Am Coll Radiol. 2024 Mar 16:S1546-1440(24)00292-8. doi: 10.1016/j.jacr.2024.02.034. Epub ahead of print.

77. Wenderott K, et al. Prospective effects of an artificial intelligence-based computer-aided detection system for prostate imaging on routine workflow and radiologists' outcomes. Eur J Radiol. 2024 Jan;170:111252. doi: 10.1016/j.ejrad.2023.111252. Epub 2023 Dec 6.

78. Singer, S.J., et al., Enhancing the value to users of machine learning-based clinical decision support tools: A framework for iterative, collaborative development and implementation. Health Care Manage Rev, 2022. 47(2): p. E21-e31.

79. Skivington, K., et al., A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Bmj, 2021. 374: p. n2061.

80. Brown, C.H., et al., An Overview of Research and Evaluation Designs for Dissemination and Implementation. Annu Rev Public Health, 2017. 38: p. 1-22.

81. Curran, G.M., et al., Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care, 2012. 50(3): p. 217-26.

82. Glasgow, R.E. and K.M. Emmons, How can we increase translation of research into practice? Types of evidence needed. Annu Rev Public Health, 2007. 28: p. 413-33.

83. Boulton, R., et al, The Cultural Politics of 'Implementation Science'. J Med Humanit, 2020. 41(3): p. 379-394.

84. Frank, D., et al., Demystifying theory and its use in improvement. BMJ Quality & Safety, 2015. 24(3): p. 228.

85. Rabin, K., et al., Explore D&I TMFs. 2023, Dissemination & Implementation Models in Health.

86. Maniatopoulos, G., et al., The paradoxes of implementation: a need to reframe the purpose of implementation science?, in Organizational Behaviour in Health Care (OBHC) Conference. 2022: Birmingham, UK.

87. Liberati, E.G., et al., What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci, 2017. 12(1): p. 113.

88. Nilsen, P., Making sense of implementation theories, models and frameworks. Implementation Science, 2015. 10(1): p. 53.

89. Davis, F.D., Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly, 1989. 13(3): p. 319-340.

90. Rogers, E., Diffusion of innovations. 5th ed. 2003, New York: Free Press.

91. Damschroder, L.J., et al., The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci, 2022. 17(1): p. 75.

92. Hogg, H.D.J., et al., Evaluating the translation of implementation science to clinical artificial intelligence: a bibliometric study of qualitative research. Front Health Serv, 2023. 3: p. 1161822.

93. Birken, S.A., et al., Criteria for selecting implementation science theories and frameworks: results from an international survey. Implementation Science, 2017. 12(1): p. 124.

94. Weick, K.E., Theory Construction as Disciplined Imagination. The Academy of Management Review, 1989. 14(4): p. 516-531.

95. Roman, K., Engaging with theory: from theoretically informed to theoretically informative improvement research. BMJ Quality & Safety, 2019. 28(3): p. 177.

96. Lempp, H., M. et al, The place of implementation science in the translational medicine continuum. Psychological Medicine, 2011. 41(10): p. 2015-2021.

97. Schultes, M.T., et al., Competences for implementation science: what trainees need to learn and where they learn it. Adv Health Sci Educ Theory Pract, 2021. 26(1): p. 19-35.

98. Powell, B.J., et al., A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci, 2015. 10: p. 21.

99. Fernandez, M.E., et al., Implementation Mapping: Using Intervention Mapping to Develop Implementation Strategies. Front Public Health, 2019. 7: p. 158.

100. Rapport, F. and J. Braithwaite, Are we on the cusp of a fourth research paradigm? Predicting the future for a new approach to methods-use in medical and health services research. BMC Medical Research Methodology, 2018. 18(1): p. 131.

101. The Division of Cancer Control and Population Sciences, Group-Evaluated Measures (GEM). 2023, National Cancer Institute, National Institues of Health.

102. Lee, PHA, et al., Trends in qualitative research prevalence and type in UK ophthalmology, in Royal College of Ophthalmologists Annual Congress. 2023: Birmingham, UK.

103. Shinners, L., et al., Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: An integrative review. Health Informatics J, 2020. 26(2): p. 1225-1236.

104. Young, A.T., et al., Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. The Lancet Digital Health, 2021. 3(9): p. e599-e611.

105. Ann, B., et al., Protocol for a qualitative study to explore acceptability, barriers and facilitators of the implementation of new teleophthalmology technologies between community optometry practices and hospital eye services. BMJ Open, 2022. 12(7): p. e060810.

106. Hull, L., et al., Designing high-quality implementation research: development, application, feasibility and preliminary evaluation of the implementation science research development (ImpRes) tool and guide. Implementation Science, 2019. 14(1): p. 80.

Chapter 2: Methodology

This chapter begins with personal reflections from the thesis author, to transparently share the experiences and perspectives which shaped the design and conduct of the research presented within this thesis. Relevant philosophical characteristics are then derived from these reflections and used to identify an established philosophical approach underlying this thesis. Finally, the influence of this theoretical approach over the methods selected throughout the thesis is discussed.

This chapter is intended both as an educational exercise for the author and an aid to interpretation of the thesis for its readers.

> *"While the drawing of boundaries between qualitative and quantitative is often self-protective or self-serving and unhelpful to the development of understanding, the deeper distinction between approaches to knowledge is a powerful one, and very important for seeing that there are different ways of grasping reality, or salvaging something from it."*[1]

2.1 Personal background

As a piece of research, this thesis has a practice-oriented aim to inform change to macula service delivery in the near future. The thesis' success and motivations may therefore be considered in terms of the utility assigned to it by the decision makers who initiate or influence such change.

Both the aim of this thesis and its success criteria are heavily influenced by the professional background of the lead researcher. At the outset of this research in 2021, the lead researcher had worked for 7 years as a junior hospital doctor, the most recent 5 within the specialty of ophthalmology. A key training component throughout this time was learning how to use scientific evidence to support healthcare decisions made with and for individuals in varied and complex situations. This often presented a mismatch between the relative simplicity of available evidence (e.g. for disease X, patients tend to have a better result in outcome Y with treatment Z) and the relative complexity of the reality that people experience (e.g. treatment Z requires that a specific patient spends 1 day per month away from home which, based on their understanding, is perceived as a greater cost than a poorer result in outcome Y).

This time also facilitated close observation of many senior medical and surgical colleagues at the same hospital. With similar expertise, awareness of evidence and situational awareness of clinical problems, these expert colleagues would intentionally and openly prefer (usually subtly, though not always) different management approaches. This case-based variation from clinical decision makers also applied to organisation-based variation from operational decision makers, which became apparent whilst working around different regional hospitals and hearing experiences of colleagues from more distant hospitals. It appeared that between organisations, approaches to service provision for identical clinical problems also varied greatly, even where high quality evidence might suggest better clinical or cost effectiveness from different approaches. These variations appeared largely due to factors which appeared non-clinical in nature (e.g. layout of hospital premises or leadership priorities) or were unclear.

Time spent as a practicing ophthalmologist also facilitated a close and authentic sense of the problems experienced by different stakeholders in the service. From this experience one of the largest scale and impact problems appeared to be the resource requirement of macula service provision. The challenge of meeting this resource requirement appeared to affect not just the macula service and its patients, but the whole ophthalmology service.

These personal experiences supported the development of 3 perspectives which motivated this thesis' aim, methods, and approach:

- The value of evidence is heavily influenced by the context in which it is interpreted
- Expert practitioners are led to different actions by apparently identical evidence and context
- The scientific quality of evidence does not independently determine its utility and impact for relevant stakeholders
- Improving the efficiency of macula services as soon as possible could be expected to deliver large-scale benefit to patients and services

The lead researcher also had 5 years of experience as a pre-doctoral researcher at the beginning of this research in 2021. This focused almost exclusively on quantitative methods in real-world evidence and had led to modest academic success by common measures of a junior career stage, but no perceptible influence on clinical practice or impact for the real-world patients which the evidence generated aimed to serve. This perspective served as a personal source of dissatisfaction and a motivation to understand better what forms of evidence are valued by stakeholders in healthcare and why and how evidence may come to support change. This motivated a selection of a wider variety of methods for the present thesis and in interest in the field of implementation science, with its focus on closing 'know-do gaps'.[2] These choices aimed to improve the chances of satisfying the aforementioned success criteria for the thesis and to deliver an educational experience for the lead researcher to improve the impact of their current and future work.

To facilitate the transparent translation of these personal experiences and perspectives into an overarching scientific approach for this thesis, they are explored in this chapter using high-level concepts from research philosophy.

2.2 Rationale

To support accurate and full interpretation of research evidence, it is widely considered good practice for researchers to explicitly share their perspectives regarding the nature of evidence and its means of generation.[3, 4] By understanding these perspectives, decision makers using the evidence generated by research can more fully interpret it, understanding and accommodating the limitations of its relevance for the particular decision they are trying to make. These decision makers hold various formal and informal roles with varied scope of influence. They include policy makers, healthcare managers or leaders, clinicians or even patients and other advisors they engage to make about their own care. The researcher themselves, and indeed other researchers, may also act as decision makers, e.g. deciding who to target with evidence they generate and what the nature of that evidence should be. For the researcher, the process of explicitly considering and stating their own perspectives can be a valuable reflective and educational process, ultimately improving the quality of present and future work.

This chapter does not seek to engage with the nuance of research philosophy or its taxonomy, as doing so may distract from the empirical research presented in this thesis, rather than improve its transparency and accessibility. Instead, a simple, and largely descriptive, approach has been taken to setting out underlying philosophical perspectives. This also intends to avoid what Williamson described as the increasingly "*impoverished, desiccated and confined*" way that social science research demands a researcher to:

> *"dig a hole, stick the name of this discipline or a method on it, get into it, and talk only to those who want to get into the hole with us; and they are only allowed in once they have learned the methodological rules."*[1]

2.3 Aim

This chapter aims to concisely describe the lead researcher's philosophical perspectives on the nature of truth (ontology), how it can be understood (epistemology) and how individuals' values can influence the research itself (axiology). An overarching approach for this constellation of philosophical perspectives will then be proposed and related back to the thesis.

2.4 Ontology

Ontology is the study of the nature of reality and truth.[5] This research assumes that for the phenomena under study, no truth is entirely universal. Even if this assumption is incorrect, this research assumes it to be of limited consequence as such truth is not held by any or many of the decision makers it aims to serve. It is these decision makers' perception of truth (and how that perception was, and continues to be, formed) that this research prioritises as it can inform the intentional influence of the phenomena under study.

Whilst this research does not assume that truth is valid for all individuals across all contexts, it does assume that many 'truths', particularly at a higher level of abstraction, can remain valid across a wide range of contexts. Therefore, although truth is assumed to be subjective, it is also assumed that many effectively identical truths are commonly held among different individuals, contexts, and times. Such sustained beliefs can be used as a proxy for objective truth.[6]

2.5 Epistemology

Epistemology is the study of how understanding of reality and truth can be gained.[5] This research assigns greatest value to evidence which the decision makers seeking to use it find useful and dependable for the decisions they are trying to make, e.g. if and how to implement AI-enabled macula services. It assumes that biasing evidence (i.e. the influence of the researcher on a phenomenon as they study it) is an unavoidable part of generating it, though this bias can and should be actively mitigated against through research design. The researcher should also actively seek to identify and highlight the (likely significant) residual bias in the evidence generated and acknowledge that they will at least partly fail to identify it all.

This research assigns greater value to the utility of the evidence it produces (as perceived by the decision makers it seeks to serve) than the minimisation of bias, though both are desirable. These assumptions motivated regular input from study advisory and reference groups throughout the research, to help mitigate against a broad scope of biases and to continually consider whether the value of evidence produced could be improved through changes to research design. Examples include jointly interpreting synthesised findings in chapter 3 with various stakeholder representations and separately in chapter 4 with reference group members directing the purposive sampling of an additional charity professional participant.

2.6 Axiology

Axiology is the study of the influence of the values held by researchers over the research process.[5] This research assumes that the researchers' own values, research participants' values and values held by decision makers using the evidence created, will all influence its interpretation and impact. The degree of this influence is assumed to be variable depending upon the type of evidence generated. This variability in the scope for differing interpretations of a piece of evidence is seen as desirable rather than problematic. Deciding upon the type of evidence to generate and the breadth of meaning which it may carry for different individuals is assumed to be a responsibility of the researcher. The skill in discharging this responsibility is in aligning the nature of the evidence they generate with the aim of the research they are designing and conducting.

2.7 Philosophical approach

The perspectives and aims underlying the design and conduct of the research within this thesis cannot be satisfied by fixed extreme approaches such as positivism or interpretivism.[5] Instead, they require an approach which accommodates flexibility in its view of truth and how best to come to understand it. This flexibility is applied to maximise the dependability and the utility of the evidence generated as perceived by the decision makers who seek to use it. The pragmatist approach satisfies these requirements, freeing the researcher from prescriptive definitions of truth or how it should be evidenced and enabling them to do what works.[6, 7]

It has been proposed that many clinician researchers would identify as pragmatists, both in the conversational sense of the word and as a philosophical approach.[8] Crucially, to be a pragmatist is not to ignore epistemology and ontology in order to get research done.[9] In ontological terms, pragmatists do not believe in universal truth nor that it is required. They measure the value of truth or evidence by its utility in addressing practical questions to enable action.[8] Dewey illustrated the flexible epistemological perspectives of pragmatism with his statement that, "*the ultimate end and test of all inquiry is the transformation of a problematic situation (which involves confusion and conflict) into a unified one*".[9] In axiological terms, this prioritisation of utility not only acknowledges that values held by the researcher and others could influence the research process, but demands it, as utility requires a value-based definition. Pragmatism therefore places great importance in characterising problems for research to target. As these problems are viewed as socially situated, their characterisation requires engagement with, not just the study of, the community who know the problem.[9] "The integration of particular nonexpert experience, fostered by the establishment of interaction and discussion, enables the community to better us the insights".[10]

2.8 Pragmatist influence over thesis methods

Pragmatism's focus on characterising a socially situated problem is reflected in chapters 3 and 4. In chapter 3, evidence synthesis methods and a TMF are used to identify the relevant stakeholder community and abstract the perspectives they hold about clinical AI across clinical settings. Primary qualitative methods then draw on these foundations to explore the problems experienced by macula service stakeholders and what they perceive will influence the implementation of AI-enabled macula services and why.

Pragmatism is often associated with mixed methods research, which is distinguished from multi-methods research by a synergistic, rather than purely additive, interaction of quantitative and qualitative methods applied to the same problem.[11] This characteristic of mixed methods research is demonstrated in the pivot to quantitative methods in chapter 5, which seeks to test the clinical non-inferiority of AI-enabled treatment monitoring for nAMD. In this chapter, the limitations of such forms of evidence which imply a universal truth (e.g. that there is a single 'correct' treatment decision for a given case) are put aside to permit the generation of evidence which addresses the problem described by stakeholders in chapter 4.

Chapters 6 and 7 attempt to address the response of 'so what?' that may be expected from the pragmatist. Prior chapters may evidence the potential for safe AI-enabled nAMD treatment monitoring and the factors that could influence its implementation, but their

utility is limited without understanding how that implementation might be conducted to address the problems characterised. This prompts a secondary analysis of qualitative data in chapter 6 using a separate TMF selected for its suitability to support the design of an actionable AI-enabled intervention. This analysis draws on the understanding of *what* could influence implementation and *why*, gained in chapter 4. Chapter 7 represents a further mixed methods analysis to create further evidence of how an AI-enabled care pathway for nAMD may be operationalised whilst identifying and proposing mitigations for risks. This evidence is in a form which is particularly useful to stakeholders with high influence over AI-enabled macula service implementation, e.g. regulatory Intended Use Statements (IUS).

2.9 References

1. Williams, G., Foreword, in New Qualitative Methodologies in Health and Social Care Research, F. Rapport, Editor. 2004, Routledge: London.

2. Eccles, M.P. and B.S. Mittman, Welcome to Implementation Science. Implementation Science, 2006. 1(1): p. 1.

3. Tong, A., et al, Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care, 2007. 19(6): p. 349-57.

4. Lockwood, C, et al., Chapter 2: Systematic reviews of qualitative evidence., in JBI Manual for Evidence Synthesis. 2020, Joanna Briggs Institute: Adelaide, Australia.

5. Spencer, R., et al, Philosophical approaches to qualitative research, in The Oxford Handbook of Qualitative Research, P. Leavy, Editor. 2014.

6. Peirce, C.S., What Pragmatism is. The Monist, 1905. 15(2): p. 161-181.

7. James, W., Pragmatism: A New Name for Some Old Ways of Thinking. 1907, New York and London: Longmans, Green & Co.

8. Long, K.M., et al, Being pragmatic about healthcare complexity: our experiences applying complexity theory and pragmatism to health services research. BMC Medicine, 2018. 16(1): p. 94.

9. Dillon, D.R., et al, Literacy Research in the Next Millennium: From Paradigms to Pragmatism and Practicality. Reading Research Quarterly, 2000. 35(1): p. 10-26.

10. Campbell, J., Understanding John Dewey. 1995, Chicago: Open Court. 199.

11. Cresswell, J., Research design: Qualitative, quantitative and mixed methods approaches. 3rd ed. 2008, Thousand Oaks: Sage.

Chapter 3: Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence

Problem: A meaningful qualitative evidence synthesis to inform the implementation of clinical AI has not been performed. For the present thesis, this risks research waste from duplicating unidentified work. It also limits the understanding of the stakeholder groups and perspectives which have been previously described, knowledge of which could enhance the quality of this thesis.

Objectives: This qualitative systematic review aimed to identify key stakeholders, consolidate their perspectives on clinical AI implementation and characterise theories, models and frameworks (TMFs) used in clinical artificial intelligence (AI) research.

Methods: Five databases were searched for primary qualitative studies on individuals' perspectives on any application of clinical AI worldwide (January 2014-April 2021). The language of the reports was not an exclusion criterion. Two independent reviewers performed title, abstract, and full-text screening with a third arbiter of disagreement. Two reviewers assigned the Joanna Briggs Institute (JBI) 10-point checklist for qualitative research scores for each. A single reviewer extracted free-text data relevant to clinical AI implementation, noting the stakeholders contributing to each excerpt. The best-fit framework synthesis used the Non-adoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework. To validate the data and improve accessibility, individuals representing each emergent stakeholder group codeveloped summaries of the factors most relevant to their respective groups. For the identification of TMFs the search was repeated up until October 2022. Each instance of TMF application and the way in which the TMFs were used in eligible studies was characterised.

Findings: The initial search yielded 4437 deduplicated articles, with 111 (2.5%) eligible for inclusion (median JBI 10-point checklist for qualitative research score, 8/10). Five distinct stakeholder groups emerged from the data: healthcare professionals (HCPs), patients, carers and other members of the public, developers, healthcare managers and leaders, and regulators or policy makers. All stakeholder groups independently identified a breadth of implementation factors, with each producing data that were mapped between 17 and 24 of the 27 adapted NASSS subdomains. Of 202 eligible studies in the updated search, 70 (34.7%) applied a TMF. Of the 50 TMFs applied, 40 (80%) were only applied once, with the Technology Acceptance Model applied most frequently (n=9). A minority of studies justified TMF application (n=51,58.6%).

Conclusions: Clinical AI implementation is influenced by many interdependent factors, which are in turn influenced by at least 5 distinct stakeholder groups. It appears that non-HCP stakeholder groups are currently under-represented and that TMFs are commonly not used. Future research should not only widen the representation of tools and contexts in qualitative research but also specifically investigate the perspectives of all stakeholder HCPs and emerging aspects of ML-based clinical AI implementation.

Relevance to future chapters: These data will form the basis for recruitment and data collection strategies in chapters 4. Similarly, the TMFs used in chapters 4 and 6 will be drawn from this rigorously derived library of prior approaches in clinical AI research and will be selected with a clear rationale.

3.1 Background

Clinical AI is a growing focus in academia, industry, and governments.[1-3] However, patients have benefited only in a few real-world contexts, reflecting a know-do gap called the "AI chasm".[4, 5] There is already evidence of tasks where healthcare professional (HCP) performance has been surpassed.[6] Reporting practices concerning quantitative measures of efficacy are also improving against evolving standards.[7] The rate-limiting step to patient benefit from clinical AI now seems to be real-world implementation.[8] This necessitates an understanding of how in real-world settings, each technology may interact with the various configurations of policy-, organizational-, and practice-level factors.[9, 10] Qualitative methods are best suited to produce evidence-based guidance to anticipate and manage implementation challenges; however, they remain rare in the clinical AI literature.[1, 11, 12]

Qualitative clinical AI literature has been broadly synthesized until 2013.[13] Despite accommodating the eligibility criteria, the study synthesized 16% (9/56) of qualitative studies that were eligible, prioritizing only higher-quality articles for data extraction. All the 9 studied tools were based on electronic healthcare records to support various aspects of prescribing. All except 1 of the studies were set in the United States, and all applied rule-based decision logic pre-programmed by human experts. The main findings included usability concerns for HCPs, poor integration of the data used by tools with the workflows and platforms in which they were placed, the technical immaturity of tools and their host systems, and the fact that adopters had a variable perception of the AI tools' value depending on their own experience.[13] Much of the subsequent clinical AI literature refers to ML-enabled tools, which differ from rule-based tools in ways that may limit the understanding of the clinical, social, and ethical implications of their implementation.[2] An example of such a tool is a classification algorithm that distinguishes retinal photographs containing signs of diabetic retinopathy from those that do not.[14] The tool "learned" to do this in a relatively unexplainable fashion through exposure to a great quantity of retinal imaging data accompanied by human-expert labels of whether diabetic retinopathy was present. These ML-based tools promise broader applicability and higher performance than rule-based tools that automate established human clinical reasoning methods.[2] An example of a rule-based tool is one that applies an a priori decision tree determined by human clinical experts to produce individualized management recommendations for patients.[15] Despite the differences in their mechanisms, both tool groups satisfy the OECD's definition of AI.[16]

3.2 Problem

It is unclear whether findings from the limited qualitative clinical AI evidence base is relevant to the modern focus on ML-based clinical AI.[17] Although primary qualitative research on ML-based clinical AI is growing, its pace remains relatively slow. If the impact of this important work is to be maximized, clarity is required regarding which perspectives and factors that influence implementation remain inadequately explored.[1] Insights into which stakeholder groups have and have not been represented to date and the TMFs that have been used are also needed to maximise the impact of future qualitative research.[18, 19] This is exemplified by the current thesis' need to plan an effective primary qualitative research study to understand *what* factors could influence the implementation of AI technology into macula services and *why* (chapter 4).

2.3 Rationale

Just 4 primary qualitative studies were identified across 2 recent syntheses of ML-based tools, and so it appears that broader eligibility criteria than previously used will be required to synthesize a meaningful volume of research at present.[11, 12] To deliver this, the present qualitative evidence synthesis has been designed to identify research of rule-based and ML-based clinical AI in any context or language. To maximise the value of the work, qualitative data will then be synthesised into findings regarding the factors that influence clinical AI implementation aiming to support future qualitative research of clinical AI in all contexts. To guide the design of future primary qualitative research, an abstracted list of stakeholder groups from existent qualitative clinical AI research and the TMFs that are applied will also be outputs of the synthesis.

3.4 Aim

This qualitative evidence synthesis aimed to identify key stakeholder groups in clinical AI implementation, consolidate their published perspectives and curate the TMFs that have explicitly been used in this field of research. This process of synthesising perspectives aimed to maximize the accessibility and utility of published data for practitioners to support their efforts to implement various clinical AI tools and to complement their insight into the unique context that they target (Figure 19). As a secondary aim, this synthesis aimed to improve the impact of future qualitative investigations of clinical AI implementation by recommending evidence-based research priorities and curating lists of participating stakeholder groups and TMFs from prior research.

- Research question
    - What are the perspectives of stakeholders in clinical artificial intelligence (AI) and how can they inform its implementation?
- Participants
    - Humans participating in primary research reporting free-text qualitative data
- Phenomena of interest
    - Individuals' perspectives of rule-based or ML-based clinical AI implementation
- Context
    - Research from any real-world, simulated, or hypothetical healthcare setting worldwide, published between January 1, 2014, and April 30, 2021, in any language
- Databases searched
    - Ovid-MEDLINE, EBSCO-CINAHL, ACM Digital Library, Science Citation Index-Web of Science, and Scopus

Figure 19 The research question, eligibility criteria informing a search strategy, and research databases that the search strategy was applied to on April 30, 2021 (Appendix 1).

3.5 Methods

This qualitative evidence synthesis adhered to an a priori protocol, the Joanna Briggs Institute (JBI) guidance for conduct and ENTREQ (Enhancing Transparency in Reporting the Synthesis of Qualitative research) reporting guidance.[20-22] The best-fit framework synthesis method was selected using the RETREAT (Review question-Epistemology-Time or Timescale-Resources-Expertise-Audience and Purpose-Type of data) criteria.[23, 24]

Following a review of implementation frameworks, the Nonadoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework was selected to accommodate the interacting complexity of factors and related stakeholders, which shape the implementation of healthcare technologies at the policy, organizational, and practice level.[10] The NASSS framework consists of seven domains, which categorize the factors that can influence implementation: (1) Condition, (2) Technology, (3) Value proposition, (4) Adopters, (5) Organization, (6) Wider context, (7) Embedding and adaptation over time.[10] In addition to its focus on technological innovations and its value in considering implementation factors between policy and practice levels, NASSS can be used as a determinant or evaluation framework rather than a process model, and it applies a relatively high level of theoretical abstraction.[25] This means that NASSS can readily accommodate perspectives from various stakeholders, contexts, and tools without enforcing excessive assumptions about the mechanisms of implementation, which is well-suited to the heterogeneous literature to be synthesized.[26]

3.5.1 Search strategy and selection criteria

The research question and eligibility criteria informed a pre-planned search strategy (available for all databases in the appendix) that is designed with an experienced information specialist (FRB), informed by published qualitative and clinical AI search strategies and executed in 5 databases (Figure 19).[6, 11, 13, 27, 28] The search strings were designed in Ovid-MEDLINE and translated into EBSCO-CINAHL, ACM Digital Library, Science Citation Index-Web of Science, and Scopus. The exact terms used are available in the appendix, but each string combined the same 3 distinct concepts of qualitative research, AI, and healthcare with *AND* Boolean operator terms. Differing thesaurus terms and search mechanisms between the databases demanded adaptation of the original search string, but each translation was aimed to reflect the original Ovid-MEDLINE version as closely as possible and was checked for sensitivity and specificity through pilot searches before the final execution. Studies concerning AI as a treatment, such as chatbots to provide talking therapies for mental health conditions, were not eligible as they represent an emerging minority of clinical AI applications.[29] They also evoke social and technological phenomena that are distinct from AI, providing clinical decision support, and therefore, risk diluting synthesized findings with nongeneralizable perspectives. The search strategy was reported in line with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension).[30] Search results were pooled in Endnote (version 9.3.3; Clarivate Analytics) for deduplication and uploaded to Rayyan.[31] The references of any review or protocol studies returned were manually searched before exclusion along with all eligible study references. Potentially relevant missing data identified in the full-text reviews were pursued with up to 3 emails to the corresponding authors. Examples of such data included eligible protocols published ≥1 year previously without a follow-up report of the study itself or multimethod studies that appeared to report only quantitative data. Title, abstract, and full-text screening were fully duplicated by 2 independent reviewers (MA and JH) with a third arbiter of disagreement (GM). Eligible articles without full text in English were translated using an automated digital translation service between May and June 2021 (Google Translate). The validity of this approach in systematic reviews has been tested empirically and is applied routinely in quantitative and qualitative syntheses.[32, 33]

3.5.2 Data Analysis

Characteristics and an overall JBI 10-point checklist for qualitative research score was assigned for each study and discussed by 2 reviewers (MA and JH) for 9.9% (11/111) of eligible studies.[22] The remaining 90.1% (100/111) were equally divided for the independent extraction of characteristics and assignment of the JBI 10-point checklist for qualitative research scores. These characteristics included the year and type of publication, source field and impact factor, implementation context studied, TMF application, study methods and study participant type and number. For each study referring to a TMF in the body text, the stage of the research at which it had contributed and any justification for its selection was noted. The index article for the TMFs applied in eligible reports were sourced to facilitate characterization by a single reviewer (JH) following consensus exercises with a senior implementation researcher (GM). Nilsen's 5-part taxonomy of TMF types (process models, determinant frameworks, classic theories, implementation theories and evaluation frameworks) and Liberati's taxonomy of TMFs' disciplinary roots (usability, technology acceptance, organizational theories and practice theories) were applied to characterize each TMF along with its year of publication.[25, 34]

Free-text data extraction using NVivo (Release 1.2; QSR International) was performed by a single reviewer (JH) following consensus exercises with 3 other authors (MA, GM, and FRB). Data were extracted in individual excerpts, which were determined to be continuous illustrations of a stakeholder's perspective on clinical AI. A single reviewer (JH) assigned each excerpt a JBI 3-tiered level of credibility (Figure 20) to complement the global appraisal of each study provided by the JBI 10-point checklist for qualitative research.[22]

- Unequivocal
    - Findings accompanied by an illustration that is beyond reasonable doubt and, therefore, not open to challenge
- Credible
    - Findings accompanied by an illustration lacking clear association with it and, therefore, open to challenge
- Not supported
    - When neither 1 nor 2 apply and when most notably findings are not supported by the data

Figure 20. Three-tiered Joanna Briggs Institute (JBI) credibility rating applied to each data excerpt, as described in the JBI Reviewers' Manual The systematic review of qualitative data.[22]

All perspectives relating to the phenomena of interest (Figure 19) arising from participant quotations or authors' narratives were extracted verbatim from the results and discussion sections. Each excerpt was attributed to the voice of an emergent stakeholder group and a single NASSS subdomain.[10] When the researcher (JH) extracting data felt that perspectives fell outside the NASSS subdomains, a draft subdomain was added to the framework to be later reviewed and reiterated with authors with varied perspectives as per the best-fit framework synthesis method.[28] A similar approach was applied to validate the stakeholder groupings which emerged. To permit greater granularity and meaning from the synthesis of such a large volume of data, inductive themes were also created within each NASSS subdomain. The initial data-led titles for these inductive themes were generated by

the researcher extracting the data, making initial revisions as the data extraction proceeded. This was followed by several rounds of discussion with the thesis' study advisory group to review and reiterate the inductive themes alongside their associated primary data to consolidate themes when appropriate and to maximize the accessibility and accuracy of their titles.

NASSS allows researchers to operationalize theory to find coherent sense in large and highly heterogeneous data such as those in this study. However, this may limit the accessibility of the analysis for some stakeholders, as it demands some familiarity with theoretical approaches.[35] To remove this barrier, the key implementation factors arising from the NASSS best-fit framework synthesis were delineated by their relevance for the 5 stakeholder groups that arose from the data. Individuals with lived experience of each emergent stakeholder role were then invited to coproduce a narrative summary of the factors most relevant to their role. The initial step in this process was the provision of a longer draft of findings relating to each stakeholder group's perspective by the lead reviewer (JH) before the review and initial discussion with each stakeholder representative. This included a senior consultant ophthalmologist delivering and leading local services (SJT), a senior clinical academic working in clinical AI regulation and sitting on a committee advising the national government on regulatory reform (AKD), a clinical scientist working for an international MedTech company (CJK), the founder and managing director of The Healthcare Leadership Academy (JM), and a panel of 4 members of the public experienced in supporting research (reference group). In these 5 separate co-production streams, the lead reviewer (JH) facilitated discussions with each stakeholder representative (AKD, CJK, JM, SJT, and reference group), who gave feedback to prioritize and validate the data discussed. The lead reviewer then redrafted the section for further rounds of review and feedback until an agreement was reached. This second analytical step validated the findings, increased their accessibility, and aimed to support different stakeholders' empathy for one another.

To preserve methodological rigor while pursuing broad accessibility, the results were presented for 3 levels of engagement. First, we used 5 stakeholder group narratives. Second, 63 inductive themes were distributed across the 27 subdomains of the adapted NASSS framework. The final most granular level of presentation used an internal referencing system within the Results section to link each assertion of the stakeholder group narratives with its supporting primary data and inductive theme. Notably, insights relevant to a given stakeholder group's perspective were often contributed by study participants from different stakeholder groups (Figure 21). This is demonstrated by the selected excerpts contained within the 5 stakeholder group narratives, which are all followed by a brief description of the stakeholders who contributed to the excerpt.

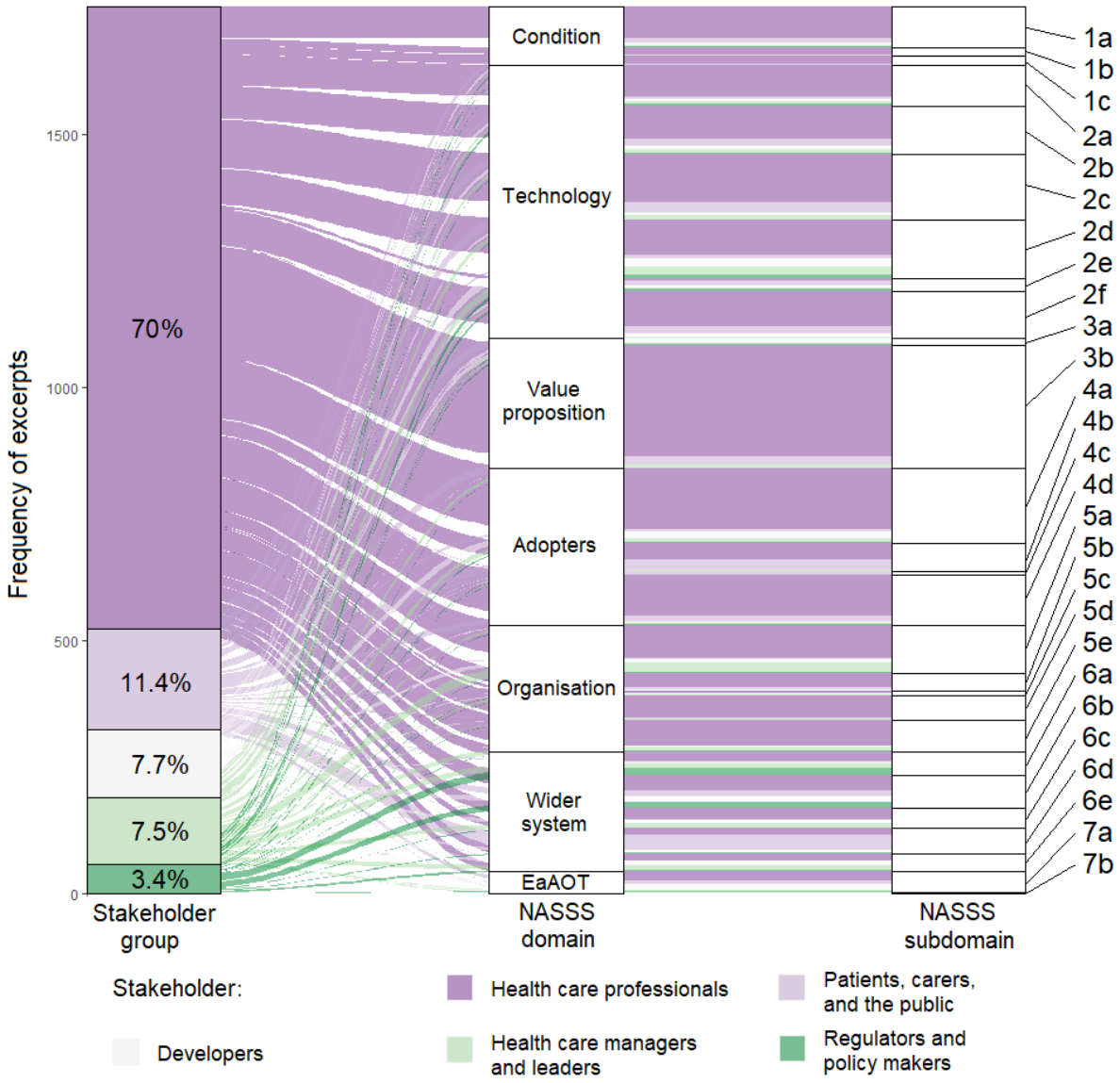


Figure 21 Sankey diagram illustrating the proportion of 1721 primary study excerpts derived from the voice of each of 5 emergent stakeholder groups and how each excerpt relates to each domain and subdomain of an adapted Non-adoption, Abandonment, Scale-up, Spread

3.6 Results

From an initial 4437 unique articles, 111 (2.5%) were found to be eligible in which 2 (1.8%) were written in languages other than English and the corresponding authors for 3 (2.7%) further studies, containing potentially relevant data, were not successfully contacted (Figure 22).[36-40]

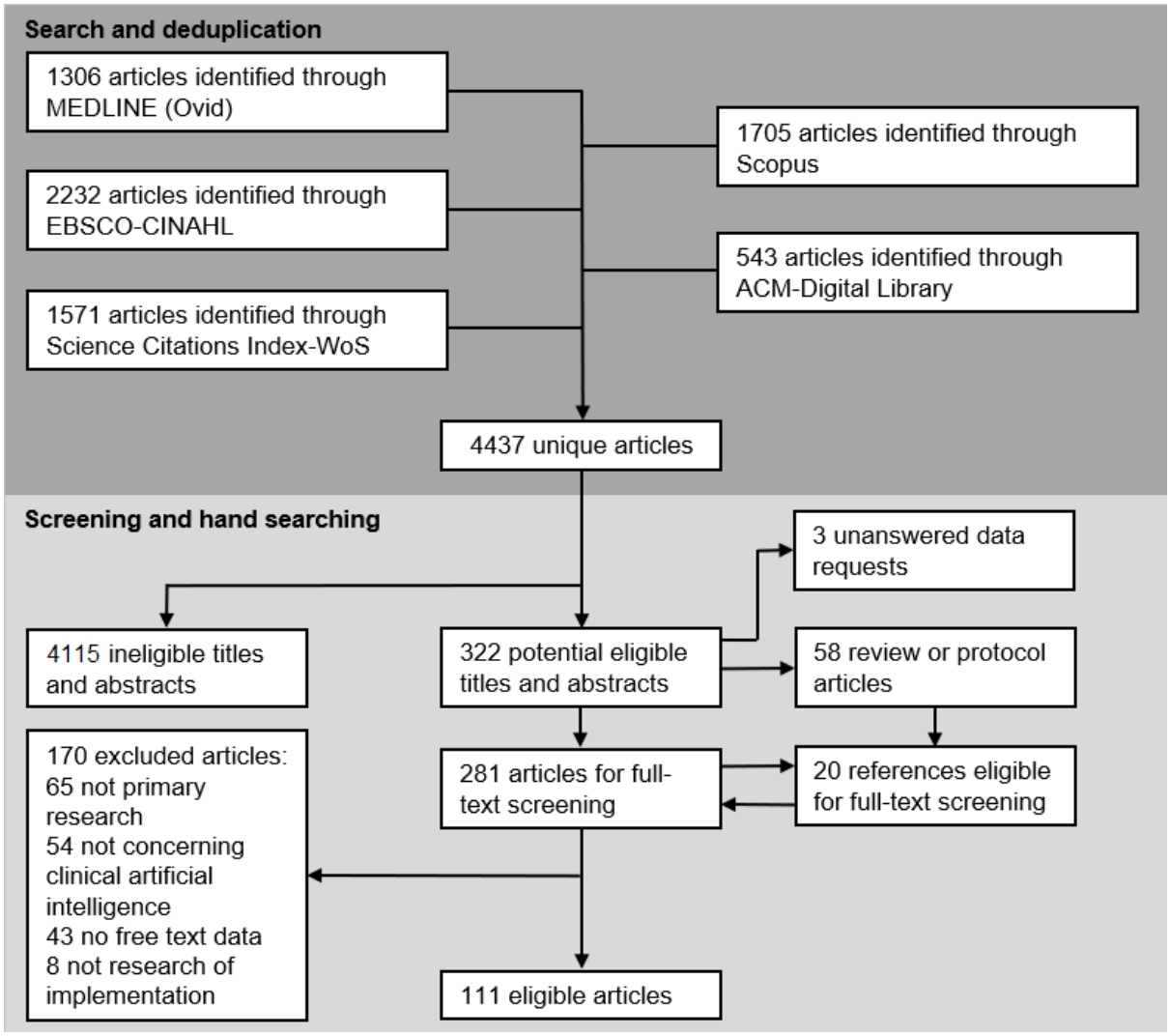


Figure 22. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) style flowchart of search and eligibility check executions [30]

Specific exclusion criteria were recorded for each excluded article at the full-text review stage, with most exclusions (4105/4326, 94.89%) made at the title and abstract screening stage. The absence of qualitative research methods was the most common cause of these exclusions. In the 111 eligible studies, there were 1721 excerpts. In assigning a JBI credibility score to each of these 1721 excerpts, 1155 (67.11%) were classified as unequivocal, 373 (21.67%) as equivocal, and 193 (11.21%) as unsupported.[22] The excerpts were categorized within the 27 subdomains of the adapted NASSS framework (Table 3) Inductive themes from within each NASSS subdomain are also listed along with the reference code applied throughout the results section and additional materials and the number of eligible primary studies which contributed.

Table 3. Subdomains of the Nonadoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework used for data analysis with 2 data-led additions to the original subdomain list.[10]

NASSS subdomain and codes	Inductive theme	Papers, n (%)

1a. Nature of condition or illness			
	1a.1	Type or format of care needs	11 (9.9)
	1a.2	Ambiguous, complicated, or rare decisions	23 (20.7)
	1a.3	Quality of current care	18 (16.2)
	1a.4	Decision urgency and impact	11 (9.9)
1b. Comorbidities			
	1b.1	Other associated health problems	5 (4.5)
	1b.2	Aligning patient and health priorities	6 (5.4)
1c Sociocultural factors		No subthemes	13 (11.7)
2a. Material properties			
	2a.1	Usability of the tool	28 (25.2)
	2a.2	Lack of emotion	12 (10.8)
	2a.3	Large amounts of changing data	14 (12.6)
2b. Knowledge to use it			
	2b.1	Knowledge required of patients	24 (21.6)
	2b.2	Enabling users to evaluate tools	20 (18)
	2b.3	Agreeing the scope of use	19 (17.1)
2c. Knowledge generated by it			
	2c.1	Communicate meaning effectively	45 (40.5)
	2c.2	Target a clinical need	23 (20.7)
	2c.3	Recommend clear action	25 (22.5)
2d. Supply model			
	2d.1	Equipment and network requirements	23 (20.7)
	2d.2	Working across multiple health data systems	25 (22.5)
	2d.3	Quality of the health data and guidelines used	33 (29.7)

2e. Who owns the intellectual property?		No subthemes	14 (12.6)
2f. Care pathway positioning[a]			
	2f.1	Extent of tools' independence	23 (20.7)
	2f.2	When and to whom the tool responds	21 (18.9)
	2f.3	How and where the tool responds	20 (18)
3a. Supply-side value (to developer)		No subthemes	7 (6.3)
3b. Demand-side value (to patient)			
	3b.1	Time required for service provision	27 (24.3)
	3b.2	Patient-centered care	22 (19.8)
	3b.3	Cost of healthcare	17 (15.3)
	3b.4	Impact on outcomes for patients	28 (25.2)
	3b.5	Educating and prompting HCPs[b]	41 (36.9)
	3b.6	Consistency and authority of care	33 (29.7)
4a. Staff (role and identity)			
	4a.1	Appetite and needs differ between staff groups	33 (29.7)
	4a.2	Tools redefine staff roles	33 (29.7)
	4a.3	Aligning with staff values	28 (25.2)
4b. Patient (simple vs complex input)			
	4b.1	Inconvenience for patients	10 (9.0)
	4b.2	Patients' control over their care	14 (12.6)
	4b.3	Aligning patients' agendas with tool use	11 (9.9)
4c. Carers		No subthemes	4 (3.6)
4d. Relationships[a]			
	4d.1	Patients' relationships with their HCPs	30 (27)
	4d.2	Users' relationships with tools	13 (11.7)

	4d.3	Relationships between health professionals	21 (18.9)
5a. Capacity to innovate in general			
	5a.1	Resources needed to deliver the benefits	29 (26.1)
	5a.2	Leadership	26 (23.4)
5b. Readiness for this technology			
	5b.1	Pressure to find a way of improving things	9 (8.1)
	5b.2	Suitability of hosts' premises and technology	15 (13.5)
5c. Nature of adoption or funding decision		No subthemes	7 (6.3)
5d. Extent of change needed to organizational routines			
	5d.1	Fitting the tool within current practices	14 (12.6)
	5d.2	Change to intensity of work for staff	22 (19.8)
5e. Work needed to plan, implement, and monitor change			
	5e.1	Training requirements	17 (15.3)
	5e.2	Effort and resources for tool launch	23 (20.7)
6a. Political or policy context			
	6a.1	Different ways to incentivize providers	10 (9)
	6a.2	Importance of government strategy	8 (7.2)
	6a.3	Policy and practice influence each other more	15 (13.5)
6b. Regulatory and legal issues			
	6b.1	Impact on patient groups	19 (17.1)
	6b.2	Product assurance	14 (12.6)
	6b.3	Deciding who is responsible	8 (7.2)
6c. Professional bodies			
	6c.1	Resistance from professional culture	20 (18)

	6c.2	Lack of understanding between professional groups	9 (8.1)
6d. Sociocultural context			
	6d.1	Culture's effect on tool acceptability	17 (15.3)
	6d.2	Public reaction to tools varies	10 (9)
6e. Interorganizational networking		No subthemes	14 (12.6)
7a. Scope for adaptation over time			
	7a.1	Normalization of technology and decreased resistance	15 (13.5)
	7a.2	Improvement of technology and its implementation	11 (9.9)
7b. Organizational resilience		No subthemes	3 (2.7)

[a]*Indicates a subdomain added to the original NASSS framework through application of the best-fit framework synthesis method.[24]*

[b]*HCP: healthcare professional.*

Five distinct stakeholder groups emerged through the analysis, each contributing excerpts related to 17 to 24 of the 27 subdomains (Figure 21). Eligible studies (Table 4) represented 23 nations, with the United States, the United Kingdom, Canada, and Australia as the most common host nations, and 25 clinical specialties, with a clear dominant contribution from primary care. Although there was some representation from resource-limited nations, 88.2% (90/102) of the studies focusing on a single nation were in countries meeting the United Nations Development Programme's definition of "very high human development" with a human development index between 0.8 and the upper limit of 1.0.[41] The median human development index of the host nations for these 101 studies was 0.929 (interquartile range (IQR) 0.926-0.944). The JBI 10-point checklist for qualitative research scores assigned to each study had a median of 8 (IQR 7-8).[22]

Table 4. Characteristics of 111 eligible studies and the clinical artificial intelligence (AI) studied.

Characteristic		Studies, n (%)
Clinical AI application		
	Hypothetical	31 (27.9)
	Simulated	24 (21.6)
	Clinical	56 (50.5)
Clinical AI nature		

	Rule based	66 (59.5)
	ML based	41 (36.9)
	NS[a]	4 (3.6)
Clinical AI audience		
	Public	5 (4.5)
	Primary care	45 (40.5)
	Secondary care	43 (38.7)
	Mixed	3 (2.7)
	NS	15 (13.5)
Clinical AI input		
	Numerical or categorical	83 (74.8)
	Imaging	9 (8.1)
	Mixed	1 (0.9)
Clinical AI task		
	Triage	15 (13.5)
	Diagnosis	15 (13.5)
	Prognosis	10 (9)
	Management	46 (41.4)
	NS	24 (21.6)
Research method		
	Interviews	54 (48.6)
	Focus groups	19 (17.1)
	Surveys	12 (10.8)
	Think aloud exercises	1 (0.9)
	Observation	1 (0.9)
	Mixed	24 (21.6)

[a]NS: not specified.

3.6.1 Developers

The developers of clinical AI required both technical and clinical expertise alongside effective interaction within the multiple professional cultures that stakeholders inhabit (6e and 6c.2). This made cross-disciplinary work a priority, but it was challenged by the immediate demands of clinical duties that limited HCPs' engagement (5a.1). State incentive systems for cross-disciplinary work had the potential to make this collaboration more attractive for developers (6a.2); nevertheless, those who independently prioritized multidisciplinary teams appeared to increase their innovations' chances of real-world utility (2c.2). The instances when HCP time had been funded by industry or academia were highly valued (4a.3):

> *...she [an IT person with a clinical background] really bridges that gap...when IT folks talk directly to the front line, sometimes there's just the language barrier there.* [Unspecified professional] [38]

To safeguard clinical AI utility, developers sometimes built flexibility in the intended use of their clinical AI tool, to accommodate variable host contexts (2a.3). This flexibility was beneficial both in terms of the clinical "reasoning" a tool applied and where and how it could be applied within different organizations' or individuals' practice (2e and 5d.1). The usability and accessibility of clinical AI often have a greater impact on adopter perceptions than their performance (2a.1 and 2b.1). There were many examples of clinical AI abandonment from adopters who had not fully understood a tool (2b.3 and 5e.1) or organizations that lacked the capacity or experience to effectively implement it (5e.2). Vendors who invested in training, troubleshooting, and implementation consultancy were often better received:

> *I've learned...that this closing the loop is what makes the sale...sometimes, we're handed a package with the implementation science done.* [Healthcare manager] [39]

The poor interoperability of different systems has inhibited clinical AI scale-up (2d.2), but it has seemed to benefit electronic healthcare record providers, whose market dominance has driven the uptake of their own clinical AI tools (3a.1). Clinical AI developed inhouse, or by third parties, seemed to be at a competitive disadvantage (2d.1). Increasing market competition and political attention may lead to software or regulatory developments that indiscriminately enhance interoperability and disrupt this strategic issue (3a.1 and 7a). Developers were also affected by defensive attitudes from healthcare organizations and patients, many of whom distrust industry with access to the data on which clinical AI's training depends (2d.3 and 2e):

> *For example, Alibaba is entering the health industry. But hospitals only allow Alibaba to access data of outpatients, not data of inpatients. They [the IT firms] cannot get the core data [continuous data of inpatients] from hospitals.* [Policy maker] [40]

3.6.2 Healthcare Professionals

The HCPs' perspectives on clinical AI varied greatly (4a.1), but they commonly perceived value from clinical AI that facilitated clinical training (3b.5), reduced simple or repetitive tasks (3b.1 and 3b.2), improved patient outcomes (3b.4), or widened individuals' scope of practice (4a.2). Despite these incentives, HCP adoption was often hampered by inadequate time to embed clinical AI in practice (5d.1), scepticism about its ability to inform clinical decisions (6c.1 and 2c.2), and uncertainty around its mechanics (2b.2). The "black box" effect associated with clinical ML prompted varied responses, with the burden of improvement placed on either the HCP to educate themselves or developers to produce more familiar metrics of efficacy and interpretability (2c.1 and 2b.2):

> *"When I bring on a test, I usually know what method it is. You tell me AI, and I have conceptually no idea."... As a result, pathologists wanted to get a basic crash course in using AI...* [HCP] [42]

The HCP culture could be very influential in local clinical AI implementation (6c.1). Professional hierarchies were exposed and challenged through the interplay of clinical AI and professional roles and relationships (4d.3). Some experienced this as a "levelling-up" opportunity, favoring evidence over eminence-based medicine and nurturing more collaborative working environments (2d.3 and 3b.6). Others felt that their capabilities were being undervalued and even feared redundancy on occasion (4a.2):

> *The second benefit was the potential to use the deep learning system's result to prove their own readings to on-site doctors. Several nurses expressed frustration with their assessments being undervalued or dismissed by physicians.* [Authors' representation of HCPs] [14]

In some studies, HCPs felt that care provision improved both in terms of quality and reach (3b.1 and 3b.4). A virtuous cycle of engagement and value perception could develop, depending on where HCPs saw value and need in a given context (2c.2 and 2b.3). This was often when clinical AI aligned with familiar ways of working (5d.1), prompting or actioning things that HCPs knew but easily forgot (3b.5), and where the transfer of responsibility was gradual and HCP led (2f.1):

> *...to the physician, the algorithmic sorting constituted an extension of her own, and her experienced colleagues' expertise..."I consider it a clinical judgement, which we made when we decided upon the thresholds"...* [HCP] [43]

3.6.3 Healthcare Managers and Leaders

Strong leadership at any level within healthcare organizations supported successful implementation (5a.2). Competing clinical demands and the scale of projects had the potential to disincentivize initial resource investments and jeopardize the implementation of clinical AI (5e.2). Resources committed to the clinical AI implementation held more than their intrinsic value, as they signalled to adopters that implementation was a priority and encouraged a positive workforce attitude (5b.2). A careful selection of clinical AI tools that seem likely to ultimately relieve workforce pressure may help managers to protect investment and adopter buy-in despite excessive clinical burdens (3b.1 and 5b.1). Stepwise or cyclical implementation of clinical AI were also advocated as a means of smoothing workflow changes and minimizing distractions from active projects:

> *I think that if you keep it simple, and maybe in a structured way if you could layer it, so that you know, for 2012 we are focusing on these five issues and in 2013 we're focusing on these...over time you would introduce better prescribing.* [Primary care leader] [44]

The significant commitment required for effective implementation underlined the importance of judicious clinical AI selection and where, how, and for whom it would be applied (2f and 1a.3). A heuristic approach from managers' knowledge of their staff characteristics (e.g., age, training, and contract length) roughly informed a context-specific implementation strategy (4a.1). However, co-design with the adopters themselves better supported the alignment of local clinical AI values, staff priorities, and patient needs (4b.3 and 5d.1). There were examples of this process being rushed and heavy investments achieving little owing to misalignment of these aspects (2b.3 and 5a):

> *...due to shortage of capacity and resources in hospitals, business cases were often developed too quickly and procurements were made without adequate understanding of the problems needing to be addressed* [Authors' representation of healthcare managers] [45]

HCPs sometimes developed negative relationships with clinical AI, which limited sustainability if issues were not identified or addressed (4d.2 and 4a.1). Just as clinical AI with the flexibility to be applied to different local workflows appeared to be better received by adopters, an influential factor for implementation was healthcare managers who were prepared to be flexible about which part of workflow was targeted (2f). Clinical AI implementation often revealed pre-existent gaps between ideal and real-world care. Managers framed this as not only a problematic creation of necessary work but also helpful

evidence to justify greater resourcing from policymakers or higher leadership (6a.3 and 5a.1). The need to consider staff well-being by managers was also illustrated, as clinical AI sometimes absorbed simple aspects of clinical work, increasing the concentration of intellectually or emotionally strenuous tasks within clinician workflows (2a.2, 1a.2, and 5d.2):

> *The problem with implementing digital technologies is that all too often, we fail to recognise or support the human effort necessary to bring them into use and keep them in use.* [Authors' representation of HCPs] [46]

3.6.4 Patients, Carers, and the Public

Concerns about the impact of clinical AI on HCP-patient interactions mainly came from the fear of HCP substitution (4d.1). These concerns seemed strongest within mental health and social care contexts, which were felt to demand a "human touch" (1a.1, 1c, and 2a.2). Patient-facing clinical AI, such as chronic disease self-management tools, was well received if they operated under close HCP oversight (2f.1 and 2f.2). The use of clinical AI as an adjunct for narrow and simplistic tasks was more prevalent (2f.1 and 1a.2), aiming to liberate HCPs' attention to improve care quality or reach (3b.2). There were also examples of patient-facing clinical AI that appeared to better align patients and HCP agendas ahead of consultations, empowering patients to represent their wishes more effectively (4b.2 and 4c):

> *...It is an advantage when reliable information can be sent to the patient, because GPs [General Practitioners] often have to use time to reassure patients that have read inappropriate information from unreliable sources.* [HCP] [47]

There was little evidence of research into carers' perspectives. Available perspectives suggested that clinical AI could make healthcare decisions more transparent, helping carers to advocate for patients (4c). This could help anticipate and mitigate some of the reported patient inconveniences and anxieties associated with clinical AI (2b.1 and 4b.1):

> *One participant stated that the intervention needed to be "patient-centred". "Including patients in the design phase" and "conducting focus groups for patients" were suggested to improve implementation of the eHealth intervention.* [Unspecified participants] [48]

Public perception of clinical AI was extremely variable, and with little personal experience, it was common to draw on hesitancy (6d.2 and 6d.1):

> *...many women, who had a negative or mixed view of the effect of AI in society, were unsure of why they felt this way...* [Authors' representation of public] [49]

Popular media were often felt to play a key role in informing the public and to encourage expectations far removed from real-world healthcare (6d.1) However, in cases where clinical AI was endorsed by trusted HCPs overseeing their care, these issues did not appear problematic (6b.2).

3.6.5 Regulators and Policy Makers

There was a perceived need for ongoing regulation of clinical AI and the contexts in which they are applied. This was both in terms of how tools are deployed to new sites (2b.3 and 5f.2) and how they may evolve through everyday practice (2a.3 and 7a.2). To make this evolution safe, stakeholders identified the need for long-term multistakeholder collaboration (6e). However, the data highlighted disincentives for this way of working, suggesting that there may be a need to enforce it (6a.2 and 6c.2). Stakeholders also raised issues around generalizability and bias for the populations they served, which were context specific and could evolve over time (6b.1). Otherwise, practitioners could gradually apply clinical AI to specific settings for which it was not appropriately trained or validated (2b.3). This "use case creep" described in the data further supported the perceived need for continual monitoring and evaluation of adopters' interaction with clinical AI (6b):

> *...they reported use of the e-algo only when they were confused or had more difficult cases. They did not feel the time required to use the e-algo warranted its use in the cases they perceived as routine or simple.* [Authors' representation of HCPs] [15]

Stakeholders often felt that clinical AI increased the speed and strength of policy and practice's influence over one another (6a.3). Many appreciated its improvement of care consistency across contexts and alignment of practices with guidelines (3b.6 and 2d.3). Others criticized it as an oversimplification (6c.1). An opportunity was seen for policy development to become more dynamic and evidence based (3b.4). Some envisaged this as an automated quality improvement cycle, whereas others anticipated complete overhauls of treatment paradigms (2f.1).

> *I could easily see us going to that payer and saying, "Well, our risk model...shows your patient population is higher risk. We need to do more intervention, so we need more money."* [Healthcare manager] [39]

Anxiety over who would hold legal responsibility if clinical AI became dominant was common (6b.3). The litigative threat was even felt by individuals who avoided clinical AI use,

as HCPs feared allegations of negligence for not using clinical AI (6b.3). Neither industry nor clinical professionals felt well placed to take on legal responsibility for clinical AI outcomes because they felt they only understood part of the whole (2b.2 and 6e). This was mainly presented as an educational issue rather than a consequence of transparency and explainability concerns (2b.2). Such high-stakes uncertainties appeared likely to perpetuate resistance from stakeholders (6c.1) although some data suggested that legislation could prompt adaptation to commercial and clinical practices that would reassure individual adopters (6b.2):

> *...physicians stated that they were not prepared (would not agree?) to be held criminally responsible if a medical error was made by an AI tool.* [Authors' representation of HCPs] [50]

> *...content vendors clearly state that they do not practice medicine and therefore should not be liable...* [Authors' representation of developer] [51]

3.6.6 Theories models and frameworks

Due to the priority assigned to the synthesis of qualitative data and the time that it required, the search strategy was updated prior to the analysis of TMFs. Updating the search across all 5 databases up to October 2022 increased the overall number of de-duplicated potential eligible articles to 6,653, with 202 eligible following screening.

3.6.6.1 TMF characteristics

Seventy eligible reports (34.7%) applied at least one of 50 distinct TMFs in the main text (Table 5), 7 (14.0%) of these were new TMFs developed within the eligible article itself. Theory application was increasingly prevalent as studies focused closer toward real-world use, with studies of hypothetical, simulated or active clinical use cases applying TMFs in 26.9%, 34.8% and 42.3% of studies respectively. There was no significant difference between the frequency of TMF application before and after the start of 2021, the median year of publication (Chi squared test, p=0.17). Twelve (17.1%) of the 70 reports drawing on a TMF applied more than one (maximum 5 [52]). Of the 87 instances that a TMF was applied it originated from the fields of technology acceptance (n=36, 41.4%), practice theory (n=21, 24.1%), organizational theory (n=19, 21.8%) or usability (n=11, 12.6%) according to Liberati's taxonomy.[34] Similarly, under Nilsen's taxonomy of TMFs the purpose of each TMF applied could be classified as determinant framework (n=49, 56.3%), process model (n=18, 20.7%), classic theory (n=10, 11.5%), evaluation framework (n=9, 10.3%) or implementation theory (n=1, 1.1%).[25]

Table 5. Theories, models and frameworks applied by eligible reports. AI= Artificial Intelligence, GP = General Practitioners, PESTLE = Political, Economic, Sociological, Technological, Legal and Environmental

Theory, model or framework	Year of index publication	Liberati classification [34]	Nilsen classification[25]	Frequency of use
Awareness-to-Adherence Model[53]	1996	Practice theory	Process model	1

Behaviour change technique taxonomy[54]	2013	Technology acceptance	Evaluation framework	1
Behaviour change theory[55]	1977	Technology acceptance	Classic theory	1
Behaviour change wheel[56]	2011	Technology acceptance	Determinant framework	5
Biography of Artefact[57]	2010	Practice theory	Classic theory	1
Consolidated Framework for Implementation Research[58]	2009	Organizational theory	Determinant framework	7
Clinical adoption meta-model[59]	2014	Technology acceptance	Evaluation framework	1
Clinical performance feedback intervention theory[60]	2019	Technology acceptance	Determinant framework	1
Disruptive innovation theory[61]	1995	Organizational theory	Classic theory	1
Dual process model of reasoning[62]	2009	Technology acceptance	Classic theory	1
Expectancy-value theory[63]	2000	Technology acceptance	Classic theory	1
Fit Between Individuals Task and Technology[64]	2006	Technology acceptance	Evaluation framework	1
Flottorp framework[65]	2013	Practice theory	Determinant framework	1
Framework for designing user-centred displays of explanation[66]	2020	Usability	Determinant framework	2
Framework of patient orientation to applications of AI in healthcare[67]	2022	Practice theory	Process model	1
Goal directed design[68]	1995	Usability	Process model	1
Heuristic evaluation[69]	1990	Usability	Determinant framework	2
Human-computer trust conceptual framework[70]	2000	Usability	Process model	1
Innovation-decision process framework[71]	2013	Organizational theory	Classic theory	1
Intention to use AI Model[72]	2020	Technology acceptance	Determinant framework	1
Iterative, collaborative development and implementation framework[73]	2021	Organizational theory	Process model	1
Kano model of satisfaction[74]	1984	Usability	Determinant framework	1
Methontology[75]	1997	Usability	Process model	1
Machine learning maturity model[76]	2021	Technology acceptance	Determinant framework	1
GPs' determinants of attitude towards AI-enabled systems[77]	2022	Technology acceptance	Process model	1

Non-adoption, Abandonment, Scale-up, Spread and Sustainability[10]	2017	Organizational theory	Determinant framework	2
Normalisation process model[78]	2007	Practice theory	Process model	1
Normalisation process theory[79]	2009	Practice theory	Mixed	4
Occupational therapy intervention process model[80]	1998	Practice theory	Process model	1
PESTLE framework[81]	1967	Organizational theory	Evaluation framework	1
Positions of perceived control[37]	2015	Practice theory	Evaluation framework	1
Process-oriented model of implementation pathways[82]	2020	Technology acceptance	Process model	1
Programme sustainability assessment tool[83]	2014	Practice theory	Determinant framework	1
Rasmussen behaviour model[84]	1983	Usability	Classic theory	1
Rogers' Theory of Diffusion[85]	1962	Practice theory	Classic theory	1
Shackel model[86]	1991	Usability	Determinant framework	1
Sittig and Singh sociotechnical framework[87]	2010	Practice theory	Determinant framework	6
Strong structuration theory[88]	2007	Organizational theory	Determinant framework	1
Systems engineering for patient safety 3.0[89]	2020	Organizational theory	Determinant framework	1
Systems-Theoretic Accident and Process Analysis[90]	2011	Organizational theory	Evaluation framework	1
Technology acceptance model[91]	1989	Technology acceptance	Determinant framework	9
Theoretical domains framework[92]	2005	Technology acceptance	Mixed	3
Theoretical framing theory[93]	1999	Organizational theory	Classic theory	1
Theory of meaningful human control[94]	2018	Practice theory	Classic theory	1
Theory of planned behavior[95]	1991	Technology acceptance	Determinant framework	1
Two component model of attitude[96]	1961	Technology acceptance	Process model	1
Unified Theory of Acceptance and Use of Technology[97]	2003	Technology acceptance	Determinant framework	7
Usabilty criteria of Scapin and Bastien[98]	1997	Usability	Determinant framework	1
User-driven co-development of AI model[99]	2021	Practice theory	Process model	1

Work as done[100]	2015	Organizational theory	Classic theory	1

3.6.6.2 Justification and application of TMFs

TAM was the most frequent choice when a TMF was applied (n=9, 12.9%), but 40 (80.0%) of the TMFs were only applied once across all eligible reports. Across the 87 instances of reports explicitly applying a TMF, 4 different modes of application emerged; to inform the study or intervention design (n=9, 10.3%), to inform data collection (n=29, 33.3%), to inform data analysis (n=44, 50.6%) and to relate or disseminate findings to the literature (n=25, 28.7%). Most instances in which a report applied a TMF carried no explanation or justification (n=51, 58.6%). Five (5.7%) reports made isolated endorsement of the TMF's popularity or quality, e.g. "The sociotechnical approach has been applied widely ..."[101] Thirty-one (35.6%) outlined the alignment of the TMF and the present research question, e.g. "our findings are consistent with disruptive innovation theory…"[102] Eleven (12.6%) reports discussed the disadvantages and alternatives that had been considered, e.g. "Because this model does not consider the unique characteristics of the clinical setting… we further adopted qualitative research techniques based on the CFIR [Consolidated Framework for Implementation Research] to further identify barriers and facilitators of the AI-based CDSS [Clinical Decision Support System]."[103]

3.7 Discussion

These data highlight the breadth of the interdependent factors that influence the implementation of clinical AI. They also highlight the influence of at least 5 distinct stakeholder groups over each factor (Figure 21): developers, HCPs, healthcare managers and leaders, public stakeholders, and regulators and policy makers. It should be emphasized that most individuals belong to more than one stakeholder group simultaneously, and the clinical AI tool and context under consideration will transform the influence of any given implementation factor; thus, robust boundaries and weightings between different stakeholders are inevitably artificial. However, to provide a simplified overview, the common factors related to each stakeholder group's perspective are summarized in Table 6.

Table 6. A summary of common factors influencing clinical artificial intelligence (AI) implementation from 5 different stakeholder perspectives.

Stakeholder group	Common factors influencing clinical AI implementation
Developers	• Understanding clinical needs • Producing clinical AI tools capable of adapting to clinical and organizational changes • Safeguarding value in a dynamic and uncertain market
Healthcare professionals	• Feeling able to make sense of clinical AI tools in the context of their own practice • Accounting for changes to patient and professional relationships • Managing disruption to current care pathways

Healthcare managers and leaders	• Anticipating the resources required to enable implementation • Engaging all adopters early in implementation • Remaining reflexive and reactive throughout implementation
Patients, carers, and the public	• Understanding what clinical AI will mean for access to healthcare professionals • Gaining access into clinical decision-making • Reconciling varied perceptions and experiences of clinical AI
Regulators and policy makers	• Establishing mechanisms for the longitudinal monitoring of the clinical AI tool and implementation context • Strengthening the bidirectional influence of policy and practice • Achieving clarity over clinical and technical accountability

The strong representation of HCPs' perspectives in the literature is an asset. However, the 30.04% (517/1721) of the excerpts from all other stakeholder perspectives clearly hold important but underexplored insights across all implementation factors (Figure 21), which should be prioritized in future research. The underrepresentation of certain stakeholders is partly masked by the need to group together the least represented stakeholders to permit meaningful synthesis, exemplified by the total of 0.35% (6/1721) of excerpts, which is related to the carer perspective. Failure to reform this clinician-centricity will limit the understanding and management of the inherent multistakeholder process of implementation.

Whilst the relationship between TMF application and quality in qualitative research is not established, rationalised TMF usage is widely accepted as one of the core components of high quality implementation research.[22] As such, the finding that most eligible studies did not apply a TMF and that among those the majority shared no explanation for TMF selection suggests that future primary qualitative research of clinical AI could improve in this regard. The consolidated list of TMFs used and the goals and rationale for their use should support the design of future high quality implementation research.

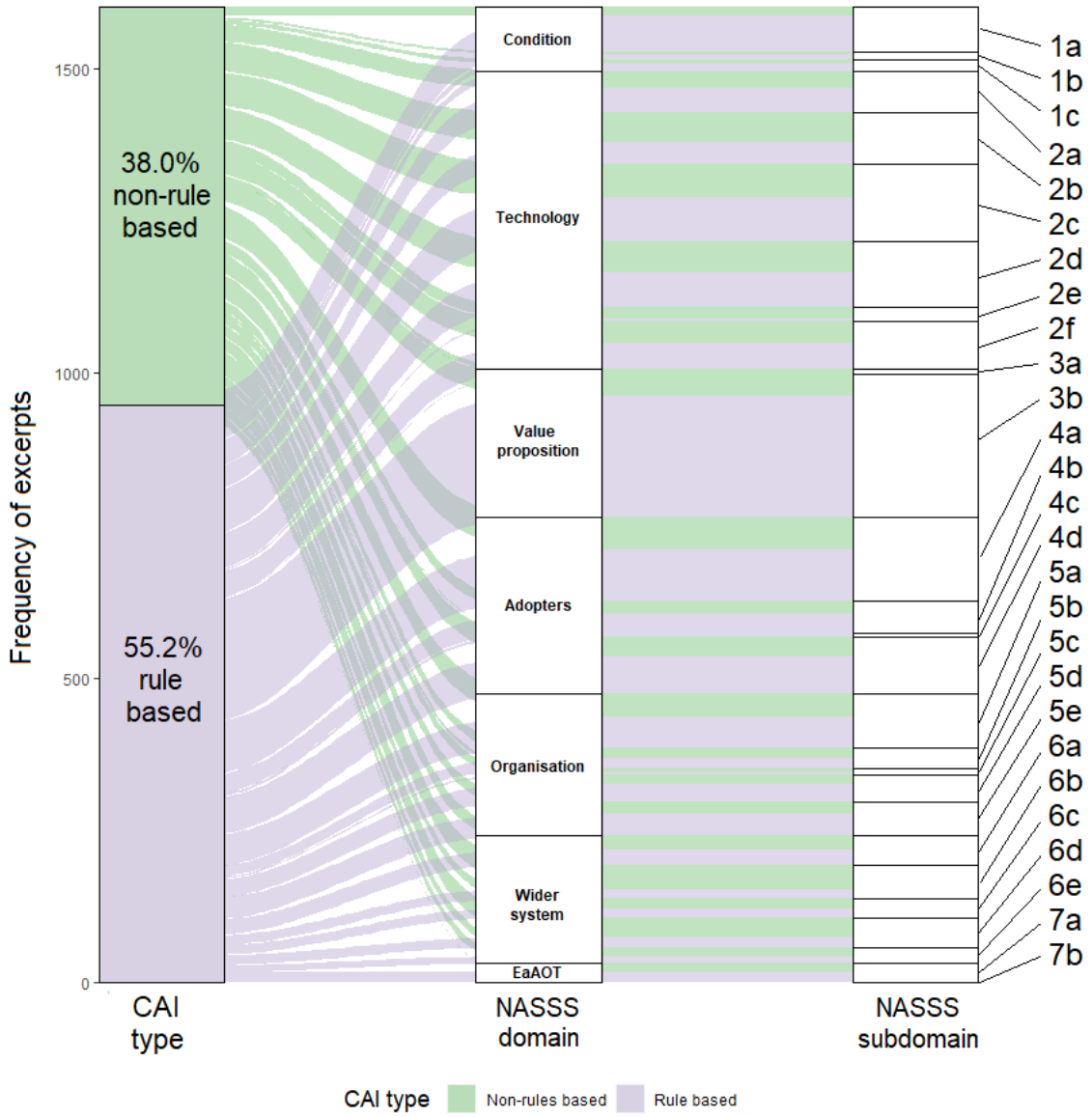


Figure 23 Sankey plot describing the relative frequency with which excerpts from eligible studies of rule-based and non-rule-based clinical artificial intelligence (CAI) relate to the various sub-domains of an adapted Non-adoption, Abandonment, Scale-up, Spread and Sustainability (NASSS) framework.

Encouragingly, the frequency at which specific factors arose in studies of rule-based and ML-based tools seemed largely comparable in Figure 23 and Table 7. This supports the use of the wider general clinical AI evidence base to inform ML-based tool implementation. Findings and methods from this evidence base have been curated and characterized here to support future technology and context-specific implementation efforts in anticipating and managing a unique constellation of factors and stakeholders. This is caveated in more dominant areas of discussion for ML-based tools, such as intellectual property, regulation, and sociocultural attitudes, where further research specific to ML-based clinical AI is required.

Table 7 The relative frequency with which excerpts from eligible studies of rule-based and Machine learning (ML)-based clinical artificial intelligence (CAI) relate to the various sub-domains of an adapted Non-adoption, Abandonment Scale-up, Spread and Sustainability (NASSS) framework.

NASSS domain	NASSS sub-domain	Excerpts by NASSS sub-domain	
		ML-based CAI	Rule-based CAI
1. Condition	1a Nature of condition or illness	21.6%	78.4%
	1b Comorbidities	30.8%	69.2%
	1c Sociocultural factors	36.8%	63.2%
2. Technology	2a Material properties	40.6%	59.4%
	2b Knowledge to use it	56.6%	43.4%
	2c Knowledge generated by it	44.1%	55.9%
	2d Supply model	47.2%	52.8%
	2e Who owns the intellectual property?	87.0%	13.0%
	2f Care pathway positioning	45.6%	54.4%
3. Value proposition	3a Supply-side value (to developer)	88.9%	11.1%
	3b Demand-side value (to patient)	15.0%	85.0%
4. Adopters	4a Staff (role, identity)	38.1%	61.9%
	4b Patient (simple v complex input)	38.5%	61.5%
	4c Carers	16.7%	83.3%
	4d Relationships	35.5%	64.5%
5. Organisation	5a Capacity to innovate in general	43.3%	56.7%
	5b Readiness for this technology	47.1%	52.9%
	5c Nature of adoption and/or funding decision	66.7%	33.3%
	5d Extent of change needed to organisational routines	34.8%	65.2%
	5e Work needed to plan, implement and monitor change	35.2%	64.8%
	6a Political/policy context	47.9%	52.1%

6. Wider system	6b Regulatory/legal issues	71.4%	28.6%
	6c Professional bodies	51.6%	48.4%
	6d Socio-cultural context	67.3%	32.7%
	6e Interorganisational networking	57.7%	42.3%
7. Embedding and adaption over time	7a Scope for adaption over time	46.7%	53.3%
	7b Organisational resilience	50.0%	50.0%

3.7.1 Comparison with Prior Work

This qualitative evidence synthesis has demonstrated that many implementation factors concerning early rule-based clinical AI tools continue to be influential.[104] However, the analysis and presentation of this work has prioritized enabling a varied readership to interpret data within their own context and experience rather than prescribing factors to be considered for a narrow range of clinical AI tools and contexts.[26, 35] As a result, this study has consolidated a wider scope of research than previous work to synthesize findings that can support future implementation practice and research, considering a wide range of clinical AI tools and contexts. This approach may compromise the depth of support offered by this study relative to other syntheses for particular clinical specialties, clinical AI types, or stakeholder groups.[11, 12] To maintain rigor while acknowledging the subjective value of eligible data, a systematic, transparent, and empirical approach has been adopted. This contrasts with narrative reviews in the literature, which provide valuable insights that draw more directly on the expertise of particular groups and collaborations but may not be easily generalized to diverse clinical AI tools.[8, 105]

3.7.2 Limitations

First, some of this study's findings are limited by the low representation of certain groups' perspectives in eligible studies, which necessitated highly abstracted definitions of key stakeholders to facilitate meaningful synthesis. In addition to the example of carers mentioned previously, employees of academic and commercial institutions were both termed "developers." A related second limitation of this study was the use of databases that focused on peer-reviewed literature. This search strategy is likely to have contributed to the low representation of non-HCP stakeholder groups, as peer-reviewed publications are a resource-intensive approach to dissemination that does not reward other stakeholders as closely as it does HCPs. Potential mitigation steps included the addition of social media or policy documents, but they were thought to be unfeasible for this study, given the extensive eligible literature returned by the broad search strategy applied.[106] Instead, a co-development step was added to the analysis process to reinforce the limited stakeholder perspectives that did arise from the search strategy with the stakeholder representatives' lived experience. This was also valuable because it helped mitigate a further source of bias from factors relevant to given stakeholders that were often being described in the primary data by participants from different stakeholder groups. This is reflected in the sources of the sample excerpts interspersing the results section and by the 61 excerpts attributed to the patient (4b) or carer (4c) NASSS subdomains, 57% (35/61) were sourced from stakeholders

outside the public, patients, and carer stakeholder group. In addition to mitigating these limitations, the co-development step of analysis was also intended to help improve the accessibility of implementation science within clinical AI, where theory-focused emphasis (dogma) often obscures the value for practitioners.[35, 107] A third limitation is the likely underrepresentation of non-English language reports of studies, despite the English language limits only being applied through database indexing. Search strings devised in other languages or searches deployed in databases that focus on non-English literature could examine this potential limitation. Finally, without clear incentives for authors to report the perceived impact, mode, or rationale of TMF application, a lack of information regarding TMF use in eligible articles does not exclude a theoretical foundation. This risk of over-interpreting negative findings is not unique to the present study but is a further limitation to hold in mind.[108]

3.7.3 Future Directions

For clinical AI implementation research in general, the relatively short list of eligible qualitative studies derived from such broad eligibility criteria emphasizes the need for more primary qualitative research to explore the growing breadth of clinical AI tools and implementation contexts. Future primary qualitative studies should prioritize the perspectives of non-HCP stakeholders. Researchers may wish to couple the relevant data curated here and a rationally selected theoretical approach to develop their sampling and data collection strategies.[109, 110] Further exploration of implementation factors more pertinent to ML-based tools, such as intellectual property, regulation, and sociocultural attitudes, may also improve the literature's contemporary relevance.

Findings from this chapter informed the study design of the primary research conducted in this thesis. Chapter 4 aims to explore *what* could influence the implementation of AI technology in macula services and *why*. The abstracted stakeholder groups derived from the present synthesis were used to inform participant recruitment. This chapter's synthesised findings have informed data collection through the design of topic guides and the NASSS framework, adjusted here specifically for clinical AI, will be used to support data analysis.

3.8 Conclusions

This study has consolidated multistakeholder perspectives of clinical AI implementation in an accessible format that can inform clinical AI development and implementation strategies involving varied tools and contexts. It also demonstrates the need for more qualitative research on clinical AI, which more adequately represents the perspectives of the many stakeholders who influence its implementation and the emerging aspects of ML-based clinical AI implementation. These findings supported the design of the recruitment, data collection and analysis strategies in a primary qualitative investigation of *what* could influence the implementation of AI technology to macula services, and *why*.

3.9 Appendix

3.9.1 Search strategy across five databases – Executed 30th April 2021

3.9.1.1 MEDLINE (OVID)

1 (ethnological research or ethnograph* or life stor* or women* stor* or social construct* or postmodern* or post-structural* or post structural* or poststructural* or post modern* or post-modern* or feminis* or interpretative or interpretive action research or cooperative inquir* or co operative inquir* or co-operative inquir* or existential or

unstructured or openended or open ended or life world or life-world or conversation analys?s or personal experience* or theoretical saturation or cluster sampl* or glaser* or participant observ* or human science or biographical method or heidegger* or colaizzi* or spiegelberg* or husserl* or foucault* or mixed method* or mixed-method*).ab,kf,kw,ti.

2 (corbin* adj2 strauss*).ab,kf,kw,ti.

3 (field adj (study or studies or research)).ab,kf,kw,ti.

4 (data adj1 saturat*).ab,kf,kw,ti.

5 ((discourse* or discurs*) adj1 analys?s).ab,kf,kw,ti.

6 (van adj (manen* or kaam*)).ab,kf,kw,ti.

7 (merleau adj ponty*).ab,kf,kw,ti.

8 ((interpretative or interpretive) adj (approach or research or data or method* or paradigm)).ab,kf,kw,ti.

9 (experiential adj (qualitative or knowledge or method*)).ab,kf,kw,ti.

10 ((lived or life) adj experience*).ab,kf,kw,ti.

11 ((theme* or thematic) adj1 (analys?s or data or synthesis or research)).ab,kf,kw,ti.

12 (account* adj1 (participant or patient* or clinician* or user* or professional* or carer* or family or stakeholder or open-ended or unstructured)).ab,kf,kw,ti.

13 (ethnonursing or phenomenol* or theoretical sampl* or observational method* or content analysis or emic or etic or hermeneutic* or semiotic*).af.

14 (narrative* adj (analys?s or synthes?s or data or research or methods or inquiry)).af.

15 (constant adj (comparative or comparison)).af.

16 (grounded adj (theor* or study or studies or research or analys?s)).af.

17 (purpos* adj sampl*).af.

18 (focus adj group*).af.

19 (qualitative adj1 (research or method* or data or study or studies or paradig* or analy*)).af.

20 Qualitative Research/ or Interview/ or Nursing Methodology Research/ or exp Diffusion of Innovation/

21 (Artificial intelligence or Boltzmann machine* or Long short-term memory or Gated recurrent unit or Rectified linear unit or Autoencoder or Backpropagation or Multilayer perceptron or Convnet or Support vector machine or Random forest or Lasso or Kernel or Elastic net* or Bayesian or Naive bayes or Genetic algorithm).ab,kf,kw,ti.

22 ((deep or convolutional or bayesian or neural or elastic) adj1 net*).ab,kf,kw,ti.

23 ((machine or deep or reinforcement or ensemble or convolutional) adj1 learning).ab,kf,kw,ti.

24 Big data/ or Decision support system, clinical/ or exp Algorithms/

25 ((algorithm* or computeri* or computer-based or computer based or machine-based or machine based or Computer assisted or Computer-assisted or Computer aided or Computer-aided or integrat* or technolog* or digital or electron*) adj3 (decision support or decision-support or decision aid or decision-aid)).ab,kf,kw,ti.

26 exp Health Occupations/ or exp Health Personnel/ or exp Persons/

27 (Perspective* adj1 (patient* or carer* or clinician* or doctor* or stakeholder* or nurse*)).ab,kf,kw,ti.

28 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 27

29 21 or 22 or 23 or 24 or 25

30 26 and 28 and 29

31 limit 30 to (humans and yr="2014 -Current")

3.9.1.2 CINAHL (EBSCO)

MH (Audiorecording or Interviews+ or “Grounded theory” or “Qualitative Studies” or “Research, Nursing” or “Focus Groups” or “Discourse Analysis” or “Content Analysis” or “Ethnographic Research” or “Ethnological Research” or “Ethnonursing Research” or “Constant Comparative Method” or Phenomenology or “Phenomenological Research” or “Implementation science” or “Usability study”) OR TI (Ethnonursing or ethnograph* or “life stor*” or “women’s stor*” or emic or etic or hermeneutic* or semiotic* or “participant observ*” or “social construct*” or postmodern* or post-structural* or “post structural*” or poststructural* or “post modern*” or post-modern* or feminis* or ((interpretative or interpretive) N1 (approach or research or data or method* or paradigm)) or “action research” or “cooperative inquir*” or “co operative inquir*” or “co-operative inquir*” or existential or experiential N1 (qualitative or knowledge or method*) or “human science” or “biographical method” or “theoretical sampl*” or glaser* or unstructured or open-ended or “open ended” or narrative* N1 (analys?s or synthes?s or data or research or method* or inquiry) or “life world” or life-world or “conversation analys?s” or “personal experience*” or “theoretical saturation” or “lived experience*” or “life experience*” or “cluster sampl*” or “observational method*” or “content analysis” or Heidegger* or Colaizzi* or Spiegelberg* or husserl* or Foucault* or van N1 manen* or van N1 kaam* or merleau N1 ponty* or Corbin* N2 strauss* or grounded N1 (theor* or study or studies or research or analys?s) strauss* N2 corbin* or data N1 saturat* or “field stud*” or “field research” or purpos* N1 sampl* or focus N1 group* or discourse* N1 analys?s or discurs* N1 analys?s or constant N1 comparative or constant N1 comparison or account* N1 (participant or patient* or clinician* or user* or professional* or carer* or family or stakeholder or open-ended or unstructured) paradigm* N1 qualitative (theme* or thematic) N1 (analys?s or data or synthesis or research) Perspective* N1 (patient* or carer* or clinician* or doctor* or stakeholder* or nurse*) or “mixed methods” or “mixed-methods”) OR AB (Ethnonursing or ethnograph* or “life stor*” or “women’s stor*” or emic or etic or hermeneutic* or semiotic* or “participant observ*” or “social construct*” or postmodern* or post-structural* or “post structural*” or poststructural* or “post modern*” or post-modern* or feminis* or ((interpretative or interpretive) N1 (approach or research or data or method* or paradigm)) or “action research” or “cooperative inquir*” or “co operative inquir*” or “co-operative inquir*” or existential or experiential N1 (qualitative or knowledge or method*) or “human science” or “biographical method” or “theoretical sampl*” or glaser* or unstructured or open-ended or “open ended” or narrative* N1 (analys?s or synthes?s or data or research or method* or inquiry) or “life world” or life-world or “conversation analys?s” or “personal experience*” or “theoretical saturation” or “lived experience*” or “life experience*” or “cluster sampl*” or “observational method*” or “content analysis” or Heidegger* or Colaizzi* or Spiegelberg* or husserl* or Foucault* or van N1 manen* or van N1 kaam* or merleau N1 ponty* or Corbin* N2 strauss* or grounded N1 (theor* or study or studies or research or analys?s) strauss* N2 corbin* or data N1 saturat* or “field stud*” or “field research” or purpos* N1 sampl* or focus N1 group* or discourse* N1 analys?s or discurs* N1 analys?s or constant N1 comparative or constant N1 comparison or account* N1 (participant or patient* or clinician* or user* or professional* or carer* or family or stakeholder or open-ended or unstructured) paradigm* N1 qualitative (theme* or thematic) N1 (analys?s or data or synthesis or research) Perspective* N1 (patient* or carer* or clinician* or doctor* or stakeholder* or nurse*) or “mixed methods” or “mixed-methods”)

AND

MH “Decision making, computer assisted+” OR TI (“Artificial intelligence” or “Boltzmann machine*” or “Long short-term memory” or “Gated recurrent unit” or “Rectified linear unit” or Autoencoder or Backpropagation or “Multilayer perceptron” or Convnet or “Support vector machine” or “Random forest” or Lasso or Kernel or Elastic net* or Bayesian or “Naive bayes” or “Genetic algorithm” or or (deep or convolutional or bayesian or neural or elastic) N3 net* or (machine or deep or reinforcement or ensemble or convolutional) N1 learning or (algorithm* or computeri* or computer-based or “computer based” or machine-based or “machine based” or “Computer assisted” or Computer-assisted” or “Computer aided” or Computer-aided or technol* or digital or electron*) N3 (“decision support” or decision-support or “decision aid” or decision-aid)) OR AB (“Artificial intelligence” or “Boltzmann machine*” or “Long short-term memory” or “Gated recurrent unit” or “Rectified linear unit” or Autoencoder or Backpropagation or “Multilayer perceptron” or Convnet or “Support vector machine” or “Random forest” or Lasso or Kernel or Elastic net* or Bayesian or “Naive bayes” or “Genetic algorithm” or or (deep or convolutional or bayesian or neural or elastic) N3 net* or (machine or deep or reinforcement or ensemble or convolutional) N1 learning or (algorithm* or computeri* or computer-based or “computer based” or machine-based or “machine based” or “Computer assisted” or Computer-assisted” or “Computer aided” or Computer-aided or technol* or digital or electron*) N3 (“decision support” or decision-support or “decision aid” or decision-aid))

AND

MH (“Health personnel+” or Patients+ or Caregivers or Family+ or “Health Manpower+”) OR SB (Biomedical or Nursing or “Allied Health” or “Health Services Administration” or “Core nursing”)

3.9.1.3 ACM Digital Library

(Title:(“Primary care” “Secondary care” “Tertiary care” Nurs* Carer* Caregiver* Health Healthcare Doctor* Nurse* Radiology Radiologist* Hospital* “General practice” Midwif* Surgery Surgeon* Ophthalm* Dermatol* Medic* Clinic* Pharma* Oncolog* Disease* “life sciences” Geriatri* Gerontol* Microbiolog* P*diatr* Rehabilitat* “social work” “social worker” “social workers” Psychiatry* Orthop* An*sthes* Patholog* Obstetric* Gyn*colog* Otorhinolaryngolog* Rheumatolog* H*matolog* Cardio* Audiolog* Urolog* Gastroenterolog* Physiotherap*)

OR

Abstract:(“Primary care” “Secondary care” “Tertiary care” Nurs* Carer* Caregiver* Health Healthcare Doctor* Nurse* Radiology Radiologist* Hospital* “General practice” Midwif* Surgery Surgeon* Ophthalm* Dermatol* Medic* Clinic* Pharma* Oncolog* Disease* “life sciences” Geriatri* Gerontol* Microbiolog* P*diatr* Rehabilitat* “social work” “social worker” “social workers” Psychiatry* Orthop* An*sthes* Patholog* Obstetric* Gyn*colog* Otorhinolaryngolog* Rheumatolog* H*matolog* Cardio* Audiolog* Urolog* Gastroenterolog* Physiotherap*))

AND

(Title:(“Artificial intelligence” “Boltzmann machine” “Long short-term memory” “Gated recurrent unit” “Rectified linear unit” Autoencoder Backpropagation “Multilayer perceptron” Convnet “Support vector machine” “Random forest” Lasso Kernel Bayesian “Naive bayes” “Genetic algorithm” “deep net” “deep network” “convolutional net” “convolutional network” “neural net” “neural network” “elastic net” “elastic network” “machine learning” “deep learning” “reinforcement learning” “ensemble learning” “convolutional learning” “computerised clinical decision support” “computerized clinical decision support” “computerised decision support” “computerized decision support”)

OR

Abstract:(“Artificial intelligence” “Boltzmann machine” “Long short-term memory” “Gated recurrent unit” “Rectified linear unit” Autoencoder Backpropagation “Multilayer perceptron” Convnet “Support vector machine” “Random forest” Lasso Kernel Bayesian “Naive bayes” “Genetic algorithm” “deep net” “deep network” “convolutional net” “convolutional network” “neural net” “neural network” “elastic net” “elastic network” “machine learning” “deep learning” “reinforcement learning” “ensemble learning” “convolutional learning” “computerised clinical decision support” “computerized clinical decision support” “computerised decision support” “computerized decision support”))

AND

(Title:(interview* thematic qualitative “nursing research methodology” Ethno* grounded “life story” hermeneutic semiotic “data saturation” “participant observation” “action research” “co-operative inquiry” existential “field study” “field studies” “field research” “biographical method” “Narrative inquiry” “Narrative analysis” “Narrative synthesis” “life world” “conversation analysis” “theoretical saturation” “lived experience” “life experience” “content analysis” “constant comparative” “discourse analysis” “discursive analysis”

heidegger* “Implementation research” “Implementation study” “Usability study” “usability research”)

OR

Abstract:(interview* thematic qualitative “nursing research methodology” Ethno* grounded “life story” hermeneutic semiotic “data saturation” “participant observation” “action research” “co-operative inquiry” existential “field study” “field studies” “field research” “biographical method” “Narrative inquiry” “Narrative analysis” “Narrative synthesis” “life world” “conversation analysis” “theoretical saturation” “lived experience” “life experience” “content analysis” “constant comparative” “discourse analysis” “discursive analysis” heidegger* “Implementation research” “Implementation study” “Usability study” “usability research”))

3.9.1.4 Scopus

(TITLE-ABS (interview* OR ((theme OR thematic) W/1 (analys?s OR data OR synthesis OR research)) OR (qualitative W/1 (research OR method* OR data OR study OR studies OR paradig* OR analys* OR result*)) OR nursing-research-methodology OR ethnograph* OR ethnonursing OR ethnological-research OR grounded-theor* OR grounded-stud* OR grounded-research OR grounded-analys?s OR life-stor* OR women's-stor* OR emic OR etic OR hermeneutic OR semiotic OR data-saturat* OR participant-observ* OR postmodern* OR post-structural* OR feminis* OR ((interpretative OR interpretive) W/1 (approach OR research OR data OR method* OR paradigm)) OR action-research OR co-operative-inquir* OR existential OR (experiential W/1 (qualitative OR knowledge OR method*)) OR field-stud* OR field-research OR human-science OR biographical-method* OR theoretical-sampl* OR purposive-sampl* OR (account* W/1 (participant OR patient* OR clinician* OR user* OR professional* OR carer* OR family OR stakeholder OR open-ended OR unstructured)) OR (narrative* W/1 (analys?s OR synthes?s OR data OR research OR methods OR inquiry)) OR life-world OR conversation-analys?s OR theoretical-saturation OR lived-experience* OR life-experience* OR cluster-sampl* OR observational-method* OR content-analysis OR constant-comparative OR discourse-analys?s OR discurs*-analys?s OR heidegger* OR colaizzi* OR spiegelberg* OR van-manen* OR van-kaam* OR merleau-ponty* OR husserl* OR foucault* OR corbin* OR strauss* OR glaser* OR (implementation W/0 (science OR study OR research)) OR (usability W/0 (study OR research)) OR mixed-methods OR (perspectiv* W/1 (patien* OR care* OR clinicia* OR docto* OR stakeholde* OR nurs*))) OR AUTHKEY (interview* OR ((theme OR thematic) W/1 (analys?s OR data OR synthesis OR research)) OR (qualitative W/1 (research OR method* OR data OR study OR studies OR paradig* OR analys* OR result*)) OR nursing-research-methodology OR ethnograph* OR ethnonursing OR ethnological-research OR grounded-theor* OR grounded-stud* OR grounded-research OR grounded-analys?s OR life-stor* OR women's-stor* OR emic OR etic OR hermeneutic OR semiotic OR data-saturat* OR participant-observ* OR postmodern* OR post-structural* OR feminis* OR ((interpretative OR interpretive) W/1 (approach OR research OR data OR method* OR paradigm)) OR action-research OR co-operative-inquir* OR existential OR (experiential W/1 (qualitative OR knowledge OR method*)) OR field-stud* OR field-research OR human-science OR biographical-method* OR theoretical-sampl* OR purposive-sampl* OR (account* W/1 (participant OR patient* OR clinician* OR user* OR professional* OR carer* OR family OR stakeholder OR open-ended OR unstructured)) OR (narrative* W/1 (analys?s OR synthes?s OR data OR research OR methods OR inquiry)) OR life-world OR conversation-analys?s OR theoretical-saturation OR lived-experience* OR life-experience* OR cluster-sampl* OR observational-method* OR content-analysis OR constant-comparative OR discourse-analys?s OR discurs*-analys?s OR heidegger* OR colaizzi* OR spiegelberg* OR van-manen* OR van-kaam* OR merleau-ponty* OR husserl* OR foucault* OR corbin* OR strauss* OR glaser* OR (implementation W/0 (science OR study OR research)) OR (usability W/0 (study OR research)) OR mixed-methods OR (perspectiv* W/1 (patien* OR care* OR clinicia* OR docto* OR stakeholde* OR nurs*))))

AND

(TITLE-ABS (artificial-intelligence OR boltzmann-machine* OR long-short-term-memory OR gated-recurrent-unit OR rectified-linear-unit OR autoencoder OR backpropagation OR multilayer-perceptron OR convnet OR support-vector-machine OR random-forest OR lasso OR kernel OR elastic-net* OR bayesian OR naïve-bayes OR genetic-algorithm OR ((deep OR convolutional OR bayesian OR neural OR elastic) W/1 net*) OR ((machine OR deep OR reinforcement OR ensemble OR convolutional) W/1 learning) OR ((algorithm* OR computeri* OR computer-based OR machine-based OR computer-assisted OR computer-aided OR technol* OR digital OR electron*) W/3 (decision-support OR decision-support OR decision-aid))) OR AUTHKEY (artificial-intelligence OR boltzmann-machine* OR long-short-term-memory OR gated-recurrent-unit OR rectified-linear-unit OR autoencoder OR backpropagation OR multilayer-perceptron OR convnet OR support-vector-machine OR random-forest OR lasso OR kernel OR elastic-net* OR bayesian OR naïve-bayes OR genetic-algorithm OR ((deep OR convolutional OR bayesian OR neural OR elastic) W/1 net*) OR ((machine OR deep OR reinforcement OR ensemble OR convolutional) W/1 learning) OR ((algorithm* OR computeri* OR computer-based OR machine-based OR computer-assisted OR computer-aided OR technol* OR digital OR electron*) W/3 (decision-support OR decision-support OR decision-aid))))

AND

(LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014)) AND (LIMIT-TO (SUBJAREA , "MEDI") OR LIMIT-TO (SUBJAREA , "HEAL") OR LIMIT-TO (SUBJAREA , "NURS") OR LIMIT-TO (SUBJAREA , "PHAR") OR LIMIT-TO (SUBJAREA , "IMMU") OR LIMIT-TO (SUBJAREA , "DENT") OR LIMIT-TO (SUBJAREA , "Undefined"))

3.9.1.5 Science Citation Index (Web of Science)

TS=("Artificial intelligence" or "Boltzmann machine*" or "Long short-term memory" or "Gated recurrent unit" or "Rectified linear unit" or Autoencoder or Backpropagation or "Multilayer perceptron" or Convnet or "Support vector machine" or "Random forest" or Lasso or Kernel or "Elastic net*" or Bayesian or "Naive bayes" or "Genetic algorithm" or ((deep or convolutional or bayesian or neural or elastic) NEAR/1 net*) or ((machine or deep or reinforcement or ensemble or convolutional) NEAR/1 learning) or ((algorithm* or computeri* or computer-based or "computer based" or machine-based or "machine based" or "Computer assisted" or Computer-assisted or "Computer aided" or Computer-aided or technol* or digital or electron*) NEAR/3 ("decision support" or decision-support or "decision aid" or decision-aid)))

AND

TS=(interview* or ((theme or thematic) NEAR/1 (analys?s or data or synthesis or research)) or (qualitative NEAR/1 (research or method* or data or study or studies or paradig* or analys* or result*)) or "nursing research methodology" or ethnograph* or ethnonursing or "ethnological research" or "grounded theor*" or "grounded stud*" or "grounded research" or "grounded analys?s" or "life stor*" or "women's stor*" or emic OR etic OR hermeneutic OR semiotic OR "data saturat*" OR "participant observ*" or postmodern* OR "post structural*" OR feminis* OR ((interpretative or interpretive) NEAR/1 (approach or research or data or method* or paradigm)) or "action research" OR "co-operative inquir*" or existential OR (experiential NEAR/1 (qualitative or knowledge or metho*)) OR "field stud*" OR "field research" or "human science" or "biographical method*" or "theoretical sampl*" or "purposive sampl*" or (account* NEAR/1 (participant or patient* or clinician* or user* or professional* or carer* or family or stakeholder or "open ended" or unstructured) or (narrative* NEAR/1 (analys?s or synthes?s or data or research or methods or inquiry)) OR "life world" OR "conversation analys?s" OR "theoretical saturation" or "lived experience*" OR "life experience*" or "cluster sampl*" or "observational metho*" or "content analysis" or "constant comparative" or "discourse analys?s" or "discurs* analys?s" or heidegger* or colaizzi* or spiegelberg* or "van manen*" or "van kaam*" or "merleau ponty*" or husserl* or foucault* or corbin* or strauss* or glaser* or (Implementation NEAR/0 (science or study or research)) or (Usability NEAR/0 (study or research)) or (Perspectiv* NEAR/1 (patien* or care* or clinicia* or docto* or stakeholde* or nurs*)) or Mixed-methods or "mixed methods"))

Results were then limited by 'Research Areas' to:

Medical Informatics OR Health Care Sciences Services OR Public, environmental, occupational health OR Neurosciences Neurology OR Psychiatry OR General Internal Medicine OR Radiology Nuclear Medicine Medical Imaging OR Pharmacology Pharmacy OR Mathematical Computational OR Biology OR Nursing OR Research Experimental Medicine OR Oncology OR Infectious Diseases OR Life Sciences Biomedicine and Other Topics OR Genetics Heredity OR Biotechnology/Applied OR Microbiology OR Rehabilitation OR Behavioural Sciences OR Surgery OR Substance Abuse OR Geriatrics Gerontology OR Pediatrics OR Microbiology OR Biomedical Social Sciences OR Parasitology OR Tropical Medicine OR Cardiovascular System Cardiology OR Endocrinology Metabolism OR Nutrition Dietetics OR Sport Sciences OR Toxicology OR Immunology OR Orthopedics OR

Anaesthesiology OR Emergency Medicine OR Obstetric Gynecology OR Respiratory System OR Gastroenterology Hepatology OR Pathology OR Physiology OR Social Work OR Urology Nephrology OR Womens Studies OR Audiology Speech Language OR Dentistry Oral Surgery Medicine OR Ophthalmology OR Otorhinolaryngology OR Anatomy Morphology OR Integrative Complementary Medicine OR Rheumatology OR Virology OR Allergy OR Dermatology OR Family Studies OR Medical Laboratory Technology OR Mycology OR Transplantation OR Hematology OR Reproductive Biology

3.9.2 Note on contributions

Two researchers accessed each of the excerpts comprising the data for this study (Jeffry Hogg (JH) and Mohaimen Al-Zubaidy (MA)). JH, Gregory Maniatopoulos (GM), Fiona Beyer (FRB), Dawn Teare (DT), and Pearse Keane (PAK) designed the study. JH, FRB, GM, and PAK developed the search strategy. JH, MA, GM, and FRB conducted the screening process. JH and MA extracted the study characteristics, and JH extracted excerpts. JH, James Talks (SJT), Alastair Denniston (AKD), Chris Kelly (CJK), and Johan Malawana (JM) and the study reference group analysed the data.

3.10 References

1. Zhang, J., et al., An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. The Lancet Digital Health, 2022. 4(4): p. e212-e213.

2. Health Ethics and Governance team, W.H.O., Ethics and governance of artificial intelligence for health, W.H. Organization, Editor. 2021.

3. Muehlematter, U.J., et al, Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health, 2021. 3(3): p. e195-e203.

4. Yin, J., et al, Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. J Med Internet Res, 2021. 23(4): p. e25759.

5. Topol, E.J., High-performance medicine: the convergence of human and artificial intelligence. Nat Med, 2019. 25(1): p. 44-56.

6. Liu, X., et al., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health, 2019. 1(6): p. e271-e297.

7. Liu, X., et al., Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med, 2020. 26(9): p. 1364-1374.

8. Marwaha, J.S., et al., Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digit Med, 2022. 5(1): p. 13.

9. Maniatopoulos, G., et al., Large-scale health system transformation in the United Kingdom. J Health Organ Manag, 2020. 34(3): p. 325-44.

10. Greenhalgh, T., et al., Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res, 2017. 19(11): p. e367.

11. Shinners, L., et al., Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: An integrative review. Health Informatics J, 2020. 26(2): p. 1225-1236.

12. Young, A.T., et al., Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. The Lancet Digital Health, 2021. 3(9): p. e599-e611.

13. Miller, A., et al., Integrating computerized clinical decision support systems into clinical work: A meta-synthesis of qualitative research. Int J Med Inform, 2015. 84(12): p. 1009-18.

14. Beede, E., et al., A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020, Association for Computing Machinery: Honolulu, HI, USA. p. 1–12.

15. Knoble, S.J. and M.R. Bhusal, Electronic diagnostic algorithms to assist mid-level health care workers in Nepal: a mixed-method exploratory study. Int J Med Inform, 2015. 84(5): p. 334-40.

16. The Organization for Economic Cooperation and Development, Recommendation of the Council on Artificial Intelligence. 2023.

17. McCoy, L.G., et al., Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol, 2022. 142: p. 252-257.

18. Hull, L., et al., Designing high-quality implementation research: development, application, feasibility and preliminary evaluation of the implementation science research development (ImpRes) tool and guide. Implementation Science, 2019. 14(1): p. 80.

19. Birken, S.A., et al., Criteria for selecting implementation science theories and frameworks: results from an international survey. Implementation Science, 2017. 12(1): p. 124.

20. Tong, A., et al., Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol, 2012. 12: p. 181.

21. Al-Zubaidy, M., et al., Stakeholder Perspectives on Clinical Decision Support Tools to Inform Clinical Artificial Intelligence Implementation: Protocol for a Framework Synthesis for Qualitative Evidence. JMIR Res Protoc, 2022. 11(4): p. e33145.

22. Lockwood, C., et al., Chapter 2: Systematic reviews of qualitative evidence., in JBI Manual for Evidence Synthesis. 2020, Joanna Briggs Institute: Adelaide, Australia.

23. Booth, A., et al., Structured methodology review identified seven (RETREAT) criteria for selecting qualitative evidence synthesis approaches. J Clin Epidemiol, 2018. 99: p. 41-52.

24. Andrew, B. and C. Christopher, How to build up the actionable knowledge base: the role of ‘best fit’ framework synthesis for studies of improvement in healthcare. BMJ Quality & Safety, 2015. 24(11): p. 700.

25. Nilsen, P., Making sense of implementation theories, models and frameworks. Implementation Science, 2015. 10(1): p. 53.

26. Kislov, R., et al., Harnessing the power of theorising in implementation science. Implement Sci, 2019. 14(1): p. 103.

27. DeJean, D., et al., Finding Qualitative Research Evidence for Health Technology Assessment. Qual Health Res, 2016. 26(10): p. 1307-17.

28. Nagendran, M., et al., Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. Bmj, 2020. 368: p. m689.

29. Cruz Rivera, S., et al., Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med, 2020. 26(9): p. 1351-1363.

30. Rethlefsen, M.L., et al., PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev, 2021. 10(1): p. 39.

31. Ouzzani, M., et al., Rayyan—a web and mobile app for systematic reviews. Systematic Reviews, 2016. 5(1): p. 210.

32. Jackson, J.L., et al., The Accuracy of Google Translate for Abstracting Data From Non-English-Language Trials for Systematic Reviews. Ann Intern Med, 2019. 171(9): p. 677-679.

33. Ng, S., et al., Identifying barriers and facilitators in the development and implementation of government-led food environment policies: a systematic review. Nutr Rev, 2022. 80(8): p. 1896-1918.

34. Liberati, E.G., et al., What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci, 2017. 12(1): p. 113.

35. Rapport, F., et al., Too much theory and not enough practice? The challenge of implementation science application in healthcare practice. J Eval Clin Pract, 2022. 28(6): p. 991-1002.

36. Rüppel, J., ["Allowing the Data to 'Speak for Themselves'" - The Classification of Mental Disorders and the Imaginary of Computational Psychiatry]. Psychiatr Prax, 2021. 48(S 01): p. S16-s20.

37. Liberati, E.G., et al., [Barriers and facilitators to the implementation of computerized decision support systems in Italian hospitals: a grounded theory study]. Recenti Prog Med, 2015. 106(4): p. 180-91.

38. Ash, J.S., et al., Clinical Decision Support for Worker Health: A Five-Site Qualitative Needs Assessment in Primary Care Settings. Appl Clin Inform, 2020. 11(4): p. 635-643.

39. Benda, N.C., et al., “How did you get to this number?” Stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study. Journal of the American Medical Informatics Association, 2020. 27(5): p. 709-716.

40. Sun, T.Q. and R. Medaglia, Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. Government Information Quarterly, 2019. 36(2): p. 368-383.

41. United Nations Development Programme, 2021/22 Human Development Report. 20222.

42. Cai, C.J., et al., "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proc. ACM Hum.-Comput. Interact., 2019. 3(CSCW): p. Article 104.

43. Torenholt, R. and H. Langstrup, Between a logic of disruption and a logic of continuation: Negotiating the legitimacy of algorithms used in automated clinical decision-making. Health (London), 2023. 27(1): p. 41-59.

44. Clyne, B., et al., A process evaluation of a cluster randomised trial to reduce potentially inappropriate prescribing in older people in primary care (OPTI-SCRIPT study). Trials, 2016. 17(1): p. 386.

45. Mozaffar, H., et al., Taxonomy of delays in the implementation of hospital computerized physician order entry and clinical decision support systems for prescribing: a longitudinal qualitative study. BMC Med Inform Decis Mak, 2016. 16: p. 25.

46. Pope, C. and J. Turnbull, Using the concept of hubots to understand the work entailed in using digital technologies in healthcare. J Health Organ Manag, 2017. 31(5): p. 556-566.

47. Van de Velde, S., et al., Development of a Tailored Intervention With Computerized Clinical Decision Support to Improve Quality of Care for Patients With Knee Osteoarthritis: Multi-Method Study. JMIR Res Protoc, 2018. 7(6): p. e154.

48. Jackson, B.D., et al, Design considerations for an eHealth decision support tool in inflammatory bowel disease self-management. Intern Med J, 2018. 48(6): p. 674-681.

49. Lennox-Chhugani, N., et al., Women's attitudes to the use of AI image readers: a case study from a national breast screening programme. BMJ Health Care Inform, 2021. 28(1).

50. Laï, M.C., et al, Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. J Transl Med, 2020. 18(1): p. 14.

51. Ash, J.S., et al., Multiple perspectives on clinical decision support: a qualitative study of fifteen clinical and vendor organizations. BMC Med Inform Decis Mak, 2015. 15: p. 35.

52. Terry, A.L., et al., Is primary health care ready for artificial intelligence? What do primary health care stakeholders say? BMC Medical Informatics and Decision Making, 2022. 22(1): p. 237.

53. Pathman, D.E., et al., The awareness-to-adherence model of the steps to clinical guideline compliance. The case of pediatric vaccine recommendations. Med Care, 1996. 34(9): p. 873-89.

54. Michie, S., et al., The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med, 2013. 46(1): p. 81-95.

55. Bandura, A., Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 1977. 84: p. 191-215.

56. Michie, S., et al, The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implementation Science, 2011. 6(1): p. 42.

57. Pollock, N. and R. Williams, e-Infrastructures: How Do We Know and Understand Them? Strategic Ethnography and the Biography of Artefacts. Computer Supported Cooperative Work (CSCW), 2010. 19(6): p. 521-556.

58. Damschroder, L.J., et al., Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implementation Science, 2009. 4(1): p. 50.

59. Price, M. and F. Lau, The clinical adoption meta-model: a temporal meta-model describing the clinical adoption of health information systems. BMC Medical Informatics and Decision Making, 2014. 14(1): p. 43.

60. Brown, B., et al., Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. Implementation Science, 2019. 14(1): p. 40.

61. Bower, J.L. and C.C. M., Disruptive Technologies: Catchin the Wave. Harvard Business Review, 1995. Jan-Feb: p. 11.

62. Croskerry, P., Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Adv Health Sci Educ Theory Pract, 2009. 14 Suppl 1: p. 27-35.

63. Wigfield, A. and J.S. Eccles, Expectancy–Value Theory of Achievement Motivation. Contemporary Educational Psychology, 2000. 25(1): p. 68-81.

64. Ammenwerth, E., et al, IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. BMC Medical Informatics and Decision Making, 2006. 6(1): p. 3.

65. Flottorp, S.A., et al., A checklist for identifying determinants of practice: A systematic review and synthesis of frameworks and taxonomies of factors that prevent or enable improvements in healthcare professional practice. Implementation Science, 2013. 8(1): p. 35.

66. Barda, A.J., et al, A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. BMC Medical Informatics and Decision Making, 2020. 20(1): p. 257.

67. Richardson, J.P., et al., A framework for examining patient attitudes regarding applications of artificial intelligence in healthcare. Digit Health, 2022. 8: p. 20552076221089084.

68. Cooper, A., About Face: The Essentials of User Interface Design. 1st ed, ed. I. John Wiley & Sons. 1995, USA.

69. Nielsen, J. and A.R. Molich, Heuristic evaluation of user interfaces, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1990: Seattle, Washington, USA. p. 249–256.

70. Madsen, M. and S. Gregor, Measuring human-computer trust. 2000.

71. Battilana, J. and T. Casciaro, The Network Secrets of Great Change Agents. Harvard Business Review, 2013. Jul-Aug.

72. Esmaeilzadeh, P., Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak, 2020. 20(1): p. 170.

73. Singer, S.J., et al., Enhancing the value to users of machine learning-based clinical decision support tools: A framework for iterative, collaborative development and implementation. Health Care Manage Rev, 2022. 47(2): p. E21-e31.

74. Kano, N., et al., Attractive Quality and Must-Be Quality. The Journal of the Japanese Society for Quality Control, 1984. 14: p. 39-48.

75. Fernández-López, M., et al, Methontology: from ontological art towards ontological engineering, in Proc. Symposium on Ontological Engineering of AAAI. 1997.

76. Pumplun, L., et al., Adoption of Machine Learning Systems for Medical Diagnostics in Clinics: Qualitative Interview Study. J Med Internet Res, 2021. 23(10): p. e29301.

77. Buck, C., et al., General Practitioners' Attitudes Toward Artificial Intelligence-Enabled Systems: Interview Study. J Med Internet Res, 2022. 24(1): p. e28916.

78. May, C., et al., Understanding the implementation of complex interventions in health care: the normalization process model. BMC Health Serv Res, 2007. 7: p. 148.

79. May, C.R., et al., Development of a theory of implementation and integration: Normalization Process Theory. Implementation Science, 2009. 4(1): p. 29.

80. Fisher, A.G., Uniting practice and theory in an occupational framework. 1998 Eleanor Clarke Slagle Lecture. Am J Occup Ther, 1998. 52(7): p. 509-21.

81. Aguilar, P.J., Scaning the Business Environment. 1967, New York: MacMillan Co.

82. Söling, S., et al., From sensitization to adoption? A qualitative study of the implementation of a digitally supported intervention for clinical decision making in polypharmacy. Implementation Science, 2020. 15(1): p. 82.

83. Luke, D.A., et al., The Program Sustainability Assessment Tool: a new instrument for public health programs. Prev Chronic Dis, 2014. 11: p. 130184.

84. Rasmussen, J., Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Transactions on Systems, Man, & Cybernetics, 1983. SMC-13: p. 257-266.

85. Rogers, E.M., Diffusion of Innovations. 1962, New York: Free Press of Glencoe.

86. Shackel, B. and S.J. Richardson, Human Factors for Informatics Usability. 1991: Cambridge University Press.

87. Sittig, D.F. and H. Singh, A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care, 2010. 19 Suppl 3(Suppl 3): p. i68-74.

88. Jack, L. and Kholeif, A.O.R., Introducing strong structuration theory for informing qualitative case studies in organization, management and accounting research. Qualitative Research in Organizations and Management: An International Journal, 2007. 2: p. 208-225.

89. Carayon, P., et al., SEIPS 3.0: Human-centered design of the patient journey for patient safety. Appl Ergon, 2020. 84: p. 103033.

90. Leveson, N.G., Applying systems thinking to analyze and learn from events. Safety Science, 2011. 49: p. 55-64.

91. Davis, F.D., Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q., 1989. 13(3): p. 319–340.

92. Michie, S., et al., Making psychological theory useful for implementing evidence based practice: a consensus approach. Qual Saf Health Care, 2005. 14(1): p. 26-33.

93. Klein, H.K. and M.D. Myers, A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. MIS Q., 1999. 23: p. 67-94.

94. Santoni de Sio, F. and J. van den Hoven, Meaningful Human Control over Autonomous Systems: A Philosophical Account. Front Robot AI, 2018. 5: p. 15.

95. Ajzen, I., The theory of planned behavior. Organizational Behavior and Human Decision Processes, 1991. 50(2): p. 179-211.

96. Rosenberg, M.J. and C.I. Hovland, "Cognitive, Affective and Behavioral Components of Attitudes, in Attitude Organization and Change: An Analysis of Consistency Among Attitude Components, Y.U. Press, Editor. 1960: New Haven.

97. Venkatesh, V., et al., User acceptance of information technology: Toward a unified view. MIS Quarterly, 2003. 27: p. 425-478.

98. Scapin, D.L. and J.M.C. Bastien, Ergonomic criteria for evaluating the ergonomic quality of interactive systems. Behaviour & Information Technology, 1997. 16(4-5): p. 220-231.

99. Schneider-Kamp, A., The Potential of AI in Care Optimization: Insights from the User-Driven Co-Development of a Care Integration System. Inquiry, 2021. 58: p. 469580211017992.

100. Wears, R.L., et al. The resilience of everyday clinical work. 2015.

101. Cresswell, K., et al., NHS Scotland's Decision Support Platform: a formative qualitative evaluation. BMJ Health & Care Informatics, 2019. 26(1): p. e100022.

102. Morgenstern, J.D., et al., "AI's gonna have an impact on everything in society, so it has to have an impact on public health": a fundamental qualitative descriptive study of the implications of artificial intelligence for public health. BMC Public Health, 2021. 21(1): p. 40.

103. Fujimori, R., et al., Acceptance, Barriers, and Facilitators to Implementing Artificial Intelligence-Based Decision Support Systems in Emergency Departments: Quantitative and Qualitative Evaluation. JMIR Form Res, 2022. 6(6): p. e36501.

104. Kawamoto, K., et al., Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj, 2005. 330(7494): p. 765.

105. Daye, D., et al., Implementation of Clinical Artificial Intelligence in Radiology: Who Decides and How? Radiology, 2022. 305(3): p. 555-563.

106. Ampt, E. and T. Ruiz, Workshop Synthesis: Use of social media, social networks and qualitative approaches as innovative ways to collect and enrich travel data. Transportation Research Procedia, 2018. 32: p. 93-98.

107. Boulton, R., et al, The Cultural Politics of 'Implementation Science'. J Med Humanit, 2020. 41(3): p. 379-394.

108. Sendak, M.P., et al., Looking for clinician involvement under the wrong lamp post: The need for collaboration measures. Journal of the American Medical Informatics Association, 2021. 28(11): p. 2541-2542.

109. Rabin, K., et al., Explore D&I TMFs. 2023, Dissemination & Implementation Models in Health.

110. Rabin, K., et al., Select strategies. 2023, Dissemination & Implementation Models in Health.

Chapter 4: Exploring stakeholder perspectives on AI-enabled macula services

Problem: AI-enabled nAMD care has not been clinically deployed or silently trialled anywhere before and qualitative ophthalmic AI research is yet to consider any aspect of nAMD care. When considering prior qualitative research in ophthalmic AI or nAMD care, only perspectives from patient and clinician stakeholder groups are published.

Objectives: This primary qualitative research seeks to explore perspectives from the full network of stakeholders in potential AI-enabled macula services to understand *what* is likely to affect its implementation into clinical practice and *why*. These insights will inform subsequent investigations to form recommendations on *how* (chapter 6) implementation should be conducted.

Methods: Thirty-six participants completed semi-structured interviews with a single interviewer. Participants included patients, carers, clinicians, industry, commissioners, managers, policy, charity sector and regulatory professionals. Data analysis was informed by a large multidisciplinary team including lay representatives and made use of NASSS. The version of the framework used also had minor adjustments made from the systematic review.

Findings: Through NASSS, determinants were categorised in 7 distinct domains. Condition: visual outcomes are broadly accepted as the key measure of success in nAMD care. Comorbidities are common within the nAMD population and must continue to be accommodated in nAMD care provision. Technology: the viability of the AI-enabled intervention will rely on complimentary technologies to enable interoperability with the host digital infrastructure and clinical pathways. Value proposition: the value proposition for users of the tool is broad. The financial value will primarily be experienced by the AI vendor and the macula service provider Adopters; few if any of the adopters are familiar with AI technologies and will be made more receptive if permitted the time and information required to build trust with the technology Organisation: Improving capacity for nAMD care at local, regional and national levels is a priority and clinical champions will be particularly influential of if and how that is achieved. Some key aspects of organisational readiness appear low at present. Wider system: There is limited communication and between relevant stakeholder groups which may adversely affect intervention design and implementation. Embedding and adaption over time: the need to improve capacity within nAMD care appears likely to be sustained or expanding. The context is currently also very supportive of AI innovation, but this may not be sustained.

Conclusions: There are many enablers for AI-enabled macula services and beyond proof of performance few of the remaining barriers depend on the AI technology itself. Successful implementation will depend on aligning and addressing the values of key stakeholders through a multi-faceted intervention.

Relevance to future chapters: Chapter 5 will test the ability of a candidate AI technology to meet stakeholders' minimum requirements to maintain visual outcomes in nAMD care. To build on the present chapter's insights into what could influence the implementation of AI-enabled macula services and why, chapter 6 will apply a TMF to data collected here to propose an evidence-based AI-enabled intervention for nAMD treatment monitoring.

4.1 Background

In chapter 3 multi-stakeholder perspectives across clinical AI and use cases were synthesised. This highlighted the importance of early and broad stakeholder involvement in the development of clinical AI products, interventions and clinical pathways. Unfortunately, it also highlighted the paucity of published stakeholder perspectives on AI-enabled ophthalmology. These examples either look across ophthalmology generally or at narrow use cases in diabetic retinopathy, distinct from the nAMD context.[1-4] In line with general weaknesses in the literature, all 4 of these studies drew exclusively on the perspectives of clinicians. Nevertheless, perspectives were raised which highlight factors likely to influence the implementation of AI-enabled treatment pathways for nAMD:

- Retinal imaging taken in the real-world setting, introducing previously imaging artefacts previously unencountered by the AI.[1]
- The general receptivity of ophthalmologists for AI-enabled innovation in their practice.[3, 4]
- The challenges of inter-professional communication and shared responsibility that will arise if clinical AI is used as a vehicle to decentralise care.[1, 2]

Considering the perspectives of stakeholders in nAMD care generally, there is an ample body of qualitative research, some of which draws on data from the UK .[5-7] The focus here is on patient perspectives, though clinician and carer perspectives also have some representation.[5-7] Qualitative research of perspectives from other stakeholder groups in nAMD care was not readily identifiable. This is understandable but does present some limitations to the interpretation of patient perspectives. These qualitative studies highlight other factors likely to influence the implementation of AI-enabled macula services:

- Patients' willingness to accept various inconveniences and compromises to optimise their visual outcomes.[6]
- The suboptimal state of current consultation quality with clinicians.[5, 8]
- The financial, social and logistical challenges posed by long distances to centralised ophthalmology units.[9]

The methods that have been used to construct meaning from these perspectives are also extremely variable. The challenge posed by the kind of multi-stakeholder study which chapter 3 recommends is distilling meaningful findings from a large and diverse dataset. As discussed in chapter 1, as in other fields, implementation science places an emphasis on the use of TMFs to produce actionable meaning from complex data.[10] As we have seen in chapter 3, there is little consensus over what TMFs should be used and for what purpose. It is clear however, that an approach to TMF selection which is firmly based in the study's aim and considers a wide range of TMFs rather than a sample restricted by convenience is most likely to capture the value of theory-informed research.[11] It is also clear from expert opinion and case studies that limiting oneself to a single TMF within a study may also limit the value on offer.

4.2 Problem

The relevance of qualitative data to a particular implementation effort is variable. Ideally, the implementation effort under consideration and the phenomena that participants are asked about would fall directly within their scope of experience.[12] A full anticipation of

the factors that influence a particular implementation effort is also more likely when a full breadth of stakeholders to that implementation niche are engaged.[13]

For the present effort to implement AI-enabled macula services in the NHS this presents several problems:

- AI-enabled macula services have not been clinically deployed or silently trialled anywhere before.[14]
- Qualitative ophthalmic AI research is yet to consider any aspect of nAMD care.
- Prior research in ophthalmic AI and nAMD care has engaged only patient, carer and clinician stakeholder groups.

4.3 Rationale

Given the pragmatist approach of this thesis (chapter 2) aiming to understand *how* best to implement AI-enabled macula services, the absence of closely relevant data to characterise *what* could influence its implementation and *why* must be addressed.[12] As we have seen from chapter 3 and the limited range of stakeholder groups engaged in adjacent fields of qualitative research, the primary qualitative research necessitated is likely to be most successful if it recruits a full range of stakeholder groups.[13] The impossibility of recruiting participants with lived experience of AI-enabled macula services is an inevitability of the translation stage of this innovation. Gaining insights from stakeholder perspectives remains an established and valuable form of implementation research even at this translational stage and the authenticity of participants' insights can be maximised through indirectly relevant experience.[12, 15] The lack of prior research in this specific implementation niche also demands explorative research methods, which allow a deep exploration of participants' insights and do not prescribe or restrict the nature of the data they may offer. Given the breadth of data that such an explorative, multi-stakeholder study of a hypothetical and complex care pathway is likely to yield, a TMF which helps to make sense of the resulting complexity is likely to support this thesis' pragmatic aim. This TMF should accommodate the full breadth of individual, organisational and system levels and would benefit from specific relevance to digital health or, if possible, clinical AI.

With these considerations in mind, this chapter will focus purely on exploring *what* may influence the implementation of AI-enabled macula services and *why*. In turn this will facilitate the design and application of further analysis of *how* implementation should be conducted. Semi-structured interviews were used to collect data from all the stakeholder groups of clinical AI implementation identified in chapter 3. Purposive sampling maximised the closeness of these stakeholder group representatives to future AI-enabled macula services. Data from chapter 3 informed the development of the topic guides on which semi-structured interviews were based. As a determinant framework spanning individual to system level factors, the NASSS framework (adapted specifically to clinical AI through the best-fit framework analysis of chapter 3) was selected to analyse the resulting data.[13, 16, 17]

4.4 Aim

To explore the perspectives of stakeholders of AI-enabled macula services to understand *what* is likely to affect its implementation into clinical practice and *why*. These insights will

inform subsequent investigations to form recommendations on *how* implementation should be conducted.

4.5 Methods

4.5.1 Participant sampling

Chapter 3 abstracted 5 stakeholder groups from the literature:

- Clinicians
- Public/patients/carers
- Healthcare managers/leaders
- regulators/policy makers
- industry professionals

For this qualitative study, purposive sampling aimed to represent each of these 5 stakeholder groups, understanding their high level of abstraction and pursuing opportunities to expand the granularity of each stakeholder group where it appeared valuable and feasible. Beside the lead researcher's own clinical and academic experience, the purposive sampling was also informed by study participants, who were each asked to suggest other important informants at the close of their interview. Sampling was also informed by the study's advisory and reference groups, which hold a range of public, patient, carer, clinician and academic perspectives (see appendix). These groups imparted a drive to maximise the diversity of participants. Besides satisfying this requirement for diverse representation, the cessation of novel insights arising from sequential interviews (i.e. data saturation) was used as a signal to conclude data collection.[17]

4.5.2 Data collection

Semi-structured interviews were selected over a survey tool as the evidence gap around AI macula services meant that too little was known about what questions would be important to adopt this more prescriptive approach to questioning.[18] This was felt to outweigh the potential for more generalisable findings from a larger scale of recruitment. Semi-structured interviews were also preferred over focus groups to allow more in depth and convenient explorations for participants.[19] It was also anticipated that eligible individuals may either be relatively frail and be disincentivised to engage in research away from home, or in very busy senior roles where social and political forces may influence the perspectives they wished to share in a group setting. These considerations also pointed to a 1:1 interview approach. This preference for interviews over focus groups reduced the likely scale of study recruitment and the nature of the data, which was free from direct peer influences. To some extent, these consequences have been balanced by a concluding public engagement event to this research programme discussed further in chapter 6.

For these semi-structured interviews, a choice of F2F discussion in the participants home or place of work, or videoconferencing was offered to participants. Participants were also asked if they consented for the interviews to be recorded. If so, recordings were sent to a professional transcription service where human technicians performed transcription, or an online AI-enabled transcription (Otter.ai, Mountain View, CA, USA) service where the researcher directly reviewed and corrected each transcript against the recording. These transcripts were treated confidentially and pseudonymised prior to analysis. Where participants refused for the interview to be recorded (one patient and one carer only), the

interviewer took brief notes through the interview itself and then made a voice recording in private immediately after the interview to capture the insights as fully as possible.

The questions posed in these semi-structured interviews drew on the contributions of the participants, but also on pre-prepared topic guides (see appendix). The initial topic guide was informed by the findings and theoretical framework of the systematic review of qualitative evidence (chapter 3). The lead researcher and study advisory group's preconceptions and descriptions of the implementation niche also contributed to this initial topic guide. Subsequently, the topic guide was re-iterated prior to each interview to take account of insights from emerging data and the stakeholder group which the next participant came from. This iterative approach to structuring qualitative interviews borrowed from that of realist evaluations, where understanding of the phenomena under study and the questions that are posed to participants are developed in tandem over the course of a study.[20] Following each interview, a reflective diary (see appendix) was added to, noting overall impressions of the interview and any factors perceived by the interviewer to impact the value or bias of the data, e.g. prior relationships with participants, rapport formation.[21]

4.5.3 Data analysis

The domains, subdomains and subthemes resulting from the adaption of the NASSS framework through the best-fit framework synthesis of chapter 3 was used to place the data directly into one of 63 codes.[13] It remained possible to further adapt the framework if the data appeared to demand it and there was no obligation to fill each subdomain. Data were not double coded, i.e. they were assigned to the single most relevant subtheme even when they could be related to multiple. This focused approach to coding aimed to mitigate against the scale and variation of the dataset.[22]

An iterative and participatory approach was taken to analysis, over the course of several meetings with the study's reference and advisory groups (see appendix). The primary researcher would independently review coded data, to provide narrative summaries of factors likely to influence the implementation of AI-enabled macula services. The primary researcher would then present these factors at these various meetings alongside supporting data to facilitate discussion and revision of their interpretation.

Once the narrative had been conversationally agreed, summaries of what might influence implementation and why were written for each of the 63 codes, accompanied by illustrative quotes (see appendix). With the aim of informing subsequent investigations to inform how implementation ought to be conducted, insights into factors appeared amenable to stakeholder actions we prioritised as key findings to be expressed within each of the 7 top-level NASSS domains. Implementation factors where tension appeared to exist between stakeholder groups, or voices in the study and reference advisory groups, were also prioritised for these findings.

4.6 Results

Thirty-six individuals consented to take part in 35 semi-structured interviews, with 1 of the interviews being conducted with both a patient and carer present (Table 1). Six of the participants were patients (Table 9).

*Table 8. Characteristics of non-patient participants by stakeholder group. * Did not consent to interview recording. ICB = Integrated Care Board.*

Stakeholder group	Number of participants	Characteristics
Carers	2	1 son* and 1 daughter of 2 separate patients; 1 cohabiting, 1 living nearby
Charity professionals	2	1 regional manager, 1 national director
Clinicians	13	1 retinal consultant, 2 registrar ophthalmologists, 2 advanced nurse practitioners, 2 hospital/community optometrists, 2 community optometrists, 1 GP partner, 2 ophthalmic photographers, 1 social work liaison
Commissioners	2	1 ICB clinical commissioning lead for ophthalmology, 1 ICB commissioning project manager
Industry professionals	4	2 imaging manufacturer representatives with national and international roles, 2 software manufacturer representatives with national and international roles
Managers	5	1 directorate manager, 1 outpatient clinical manager, 1 independent sector provider manager, 1 macula service administrator, 1 service innovation manager
Policy makers	1	National level policy maker
Regulatory professional	1	Senior regulatory consultant

26 interviews were conducted F2F, with 7 at the participants' homes, 18 at the participants' place of work and 1 at a public café. Nine interviews were conducted by videoconferencing software. Five individuals, two consultant ophthalmologists, a NHS trust head of IT, a professional carer and a patient's daughter who also happened to be a care home manager were invited to participate but were unable to find time or declined to do so. Fourteen patient participants consented to participate, but on further contact to set an interview date declined to participate. Notably two of these patients represented the only patients screened from non-white ethnic groups, one of whom did not speak English. The median length of interviews was 47 minutes with a range of 27 – 01:24. Findings are reported below within the 7 major domains of the NASSS framework; Condition, Technology, Value proposition, Adopters, Organisation, Wider system and Embedding and adaption over time.

*Table 9. Characteristics of patient participants. *Did not consent to interview recording*

ID	Gender	Ethnicity	Age	Miles from hospital	Professional status

1	Male	White British	70s	31	Retired business chairman
2	Female	White British	60s	5	Factory worker
3	Male	White British	80s	5	Retired teacher
4*	Female	White British	90s	13	Retired
5	Male	White British	90s	21	Retired teacher
6	Female	White British	80s	3	Retired cleaner

4.6.1 Condition

As a condition, hospital eye clinicians found nAMD relatively simple to treat and expressed frustrations over the amount of input it required from them. Despite this apparent simplicity, its management was indisputably seen as a specialist concern and primary care participants, were clear that ophthalmologists' held sole responsibility for it.

> *"Our relation to eyes is almost similar to our relation to teeth, we're better than no one, but we're not the guy." [GP]*

This was mirrored by optometrists exclusively working in primary care, for whom nAMD still presented a small minority of the presentations they dealt with. Many participants from different stakeholder groups alluded to the ownership of nAMD management by hospital eye services as a missed opportunity to free up ophthalmology capacity and make better use of a larger decentralised workforce of nurses and optometrists. Various participants felt some extension of consultant ophthalmologists' responsibility to the community could enable task-shifting to the community in nAMD care without challenging this perspective. Another incentive for participants to propose decentralisation was that frailty and co-morbidity are common among people with nAMD.

> *"They're old these people, and they've got sight difficulties. So, it's going to be quite daunting, getting on a bus and getting from the Haymarket bus station up to the RVI [hospital]."* [Community optometrist]

This participant alludes to the burden of travel for nAMD patients due to these broader elements of their health condition. Patients and hospital clinicians also mentioned challenges experienced in clinic where patients must navigate various rooms, chairs and couches. Hospital clinicians were also mindful of ocular comorbidities. They saw common incidental diagnoses of cataract or glaucoma in nAMD consultations as time-consuming yet valuable by-products of the current pathway. For patients, vision preservation was their priority, and this seemed to readily motivate them to overcome barriers to receiving treatment.

> *"…if anything [nAMD] did happen, and that [injections] was the only option, then I would just have to put that [fear of injections] way to the back of my mind, and just get on with it. Because to be perfectly truthful, the thought of not being able to see at all, would just be….really, really scary, really, really scary, yes." [Carer]*

This participant was the daughter of a patient and so was very mindful of her personal risk of nAMD, but no stakeholder group challenged visual outcome as the primary indicator of successful nAMD care.

4.6.2 Technology

All potential users of the AI technology held relatively abstract perceptions of its nature and role in their work. At a high-level, certain characteristics of the technology were consistently reported, e.g. an inability for dialogue, high performance in medical imaging interpretation and ability to improve clinical workflow efficiency.

> *"It's nice to see a person, you know? If you had this machine and then it would print out "She doesn't need any more," or, "She does." It's not like asking the doctor*" [Patient]

> *"It's no different to when you say, "Oh he needs an injection in four weeks", and when you show it to the consultant they say, "No, no, leave him." The only difference is that you can have a conversation with the consultant."* [Ophthalmology trainee]

When talking in greater detail about the AI technology, participants would commonly conflate different uses or types of AI within a single conversation, e.g. AI to detect other ocular and systemic diseases or AI that continues to learn from data it analyses rather than maintaining a fixed analytical process. The regulatory participant pointed out that such flexibility in an AI medical device would not be permissible within the approval for use it had been granted. This does highlight a risk of use case-drift from clinicians and misplaced fears and expectations from patients about AI-enabled macula services. Several participants suggested a phased implementation to address this to allow them to explore the intended use of the AI for themselves and the practical considerations they need to take when working with it. This time spend seeing the strengths and weaknesses of the AI for themselves was also expected to build trust.

> *"…as long as you're just aware of the artifacts and the errors, and all of the other things… it [AI] would be really, really useful."* [Nurse practitioner]

As this clinician who conducts consultations and injections within the service explains, this trust was not dependent on perfect performance from the AI, but their direct observation of its performance and the type and frequency of errors users could expect. Both clinicians and patients appeared more accepting of AI if they felt personally enabled to evaluate or challenge its outputs. There was a range of expectation over how much autonomy would be assigned to the tool, but by building this trust most participants seemed accepting of eventually allowing independent AI decisions to be made for at least a proportion of treatment monitoring episodes.

> *"…say that this AI technology is good enough to reliably, and with a very good true positive and very low false positive result, say that, "This patient needs to be seen. This patient could be seen in eight weeks. This patient can have injections deferred*

now." I think that would be favourably looked upon [by commissioners]." [Commissioner]

As this commissioning lead suggests, for many participants the business case for AI-enabled macula services appeared to depend on the acceptability of relatively high AI autonomy. Industry participants also highlighted that the business case would also depend on AI interoperability with an adopting organisations digital infrastructure due to the financial, technical, and clinical implications of changing these elements.

> *"If you're a site where you go, "you know what we're really bored of these Canon machines, they're not working out for us anymore. Let's move to Zeiss." It's very difficult to do that. Because you've got to break the legacy of your data and almost start all over again."* [Industry professional]

4.6.3 Value proposition

Participants described a broad potential value proposition, which went beyond simple gains in clinical effectiveness and efficiency to encompass financial, clinical and inter-organisational advantages.

One commissioning participant explained that payments made to providers of nAMD care do not respond to the way in which that care is provided. This meant that financial incentives in the current commissioning system lay with providers using AI to make efficiency savings on the staff and premises required for billable services (review appointments and injection administration), rather than expecting a higher tariff for AI-enabled care.

> *"On Saturdays you could have four consultants instead of one consultant and then juniors. So, because we just need to get the patients through, it potentially could be really expensive. But the Trust are happy to keep on going, because obviously we can't put the patients at risk."* [Manager]

This clinical manager felt the financial value proposition of improving clinic efficiency was particularly strong in the current context of paying costly staff overtime to control appointment waiting lists. Representing the organisation that stood to benefit financially, e.g. NHS Foundation Trusts, senior managers also felt that their teams would hold responsibility for AI procurement and implementation decisions for nAMD care, rather than regional commissioners. This could be challenged if a new provider, e.g. independent sector, were to make an AI-enabled proposal for newly commissioned nAMD services.

AI systems are conventionally described as 'black-boxes', indicating limited clinical explainability of their outputs.[23] However, participants most commonly attributed this characteristic to clinician-led decisions and saw AI as an opportunity to actually improve their experience of explainability. This came from patient, carers and clinician participants who shared examples of occasions when they were not clear on the current state of their disease or the rationale for certain treatment decisions.

> *"So being aware of what's going on and why and being able to discuss it is important. So what should happen when I go to the NHS is that they should do what the optician does,*

say 'Well, here it is Mr.XXX. This is what has occurred, how it's changed' and so on." [Patient]

In addition, this patient went on to illustrate the value they assigned to the explainability of clinical decisions made about him, through his choice to pay for imaging and consultation with his local optometrist prior to his hospital treatment appointment. He did this to ensure he had a clear understanding of the current state of his nAMD prior to his hospital appointments where he felt he could not be confident of the same degree of explainability. Clinicians performing injection-only clinics also reported frustrations where ambiguous documentation from colleagues disrupted their workflow, causing them to seek direct clarification from colleagues. Whilst faults were acknowledged, clinician, patient and carer participants expressed hope that AI-enabled treatment decisions could improve explainability in current management plans and also standardise the rationale underpinning those plans.

This standardisation held additional value for care pathways that required inter-organisation collaboration from community optometrists and hospital specialists.

> *"the four glaucoma consultants are not absolutely happy with the level of expertise that is available in the Sunderland area to make safe discharge into the community of their glaucoma patients... If you take away the diagnostic capability from the community optometrist but use them only as technicians and the diagnostic capability is handed over to an artificial intelligence system which has been tested and validated, then I think that would be something different."* [Commissioner]

Here a commissioning participant explains how reservations from ophthalmologists about the care quality and consistency in community optometry has prevented glaucoma service decentralisation. However, a consultant retina specialist emphasised elsewhere, that the value of this consistency relies on credible high levels of performance or monitoring procedures for the AI.

4.6.4 Adopters

For patients, the distinction between AI and other technologies from everyday life and elsewhere in healthcare was not clear. For some, there was a reluctance to engage with the detail of exactly what the AI technology was. This came from their own sense of low technology literacy and interest.

> *It wasn't that she felt like – 'oh, this is completely a game changer and I need to know lots of new things or wouldn't feel comfortable about it [AI]'. She was fairly ambivalent towards its introduction, essentially.* [Field note from unrecorded patient interview]

This reluctance only translated into an expectation to reject an AI-enabled service if there were expectations for them to engage with technologies directly. There was also a broad sense that the service's 'human touch' would not foreseeably be displaced by technology due to the requirement that a clinician (usually a nurse) gave the injection itself. A more unifying factor influencing acceptance across patient participants was the of accommodating frequent clinic attendance.

> *"So, I think that is probably the most stressful thing. It [phoning to check appointment booking] is just something that I need to do. I have found almost every time the last 10 or 11 [clinic appointments]."* [Patient]

There was clear hope from patients that an AI-enabled service could lessen this burden, not only through decentralisation as previously mentioned, but by facilitating shorter more predictable appointment lengths and increased appointment availability. A son and daughter of two separate patients were interviewed to represent carer perspectives. They did not appear to feel deeply involved with their parent's care which was in part attributed to a lack of access to accurate information and logistical challenges in attending clinics.

> *"But he [carer's brother] does ring my mum, so maybe, if my mum said, "Oh, there's this technology and you can see what I'm going through," then, he may go, "Oh, lovely, I'd love to see that. I'd love to be kept in the loop a bit." But I think, just being so far away is quite difficult."* [Carer]

For carers it seems that, with patients' consent, the opportunity to independently access AI technology outputs could enable them to feel more empowered as patient advocates.

Ophthalmologists appeared welcoming of the prospect of AI reducing the time they spent interpreting OCT rather than fearing professional displacement from it. They viewed it as a low complexity task in which they did not require much exposure to achieve and maintain clinical competence. Nursing and optometry staff also appeared broadly positive as they felt it would reduce their dependence on ophthalmology colleagues. There were some concerns that highly automated AI uses might displace them from their nAMD decision making role. This role was hard earned with additional training or competitive interviewing and offered an improvement in the variety and payment associated with their role.

> *"But, I suppose with AI, that the good thing is at least there is some research backing the fact that we have worked out what the specificity and the sensitivity of all these things are for certain conditions. However, we have trained them [junior clinicians] up by looking at a set of images and going through them together, and then we have given them feedback. But we have not actually measured how well they perform."* – [Consultant retina specialist]

This consultant's reflection on their personal risk from assuming responsibility for the unquantified quality of junior colleagues' decisions lends some legitimacy to the professional threat that allied health professionals alluded to. Surprisingly, administrative (appointments manager) and technician (photographer) staff, whose roles are likely to expand or become more important in AI-enabled services, expressed a fear that it might "*do away with my job"* [Manager] most clearly.

4.6.5 Organisation

Although operational staff (e.g. managers and IT professionals) within an adopting organisation appear set to strongly influence or make decisions around AI implementation, they (and other stakeholders) look to senior clinicians to inform these decisions. This was despite ophthalmologist participants reporting a low familiarity with the technology and limited competency in evaluating it. One industry participant highlighted that the strong influence of consultants' views can be problematic in organisations as pre-existent personal

or political issues relating to individual consultants can have disproportionate impact on implementation efforts.

> *"People distrust things and, let's face it, consultants can be a bit, "Nobody can do this as well as me." I love them dearly but…"*- [Manager]

This manager of a small team of ophthalmology consultants in an independent sector practice highlighted the challenges of trying to innovate a service without the support of consultants. Several professional stakeholders also pointed to the availability of adequately trained IT professionals as a key enabler for an organisation to manage the risks and technical challenges associated with providing AI-enabled care. Unfortunately, such NHS professionals were often felt to be absent or inaccessible to other professional groups in the NHS.

> *"…you end up with a couple of groups of people in NHS IT. You end up with people who are very passionate… [and] you find people who might apply on the regular for job openings at, let's say, Cerner or Deloitte, but not quite making the cut. That workforce mix doesn't always lend itself to being the most efficient at implementing the types of things we've been talking about"* [Industry professional]

This was echoed by another industry participant who presumed that third party management consultants would be required if NHS organisations were to successfully implement AI-enabled care. The costs required to overcome this limiting aspect of organisational readiness would be a clear disincentive, but managers and commissioners all pointed to improving the productivity of individual clinicians and nAMD services as a strategic priority.

> *"I think if any Trust anywhere in the country says that they are managing it [nAMD care] successfully they are not telling the truth, because almost all of them are firefighting…"* [Commissioner]

This commissioner also volunteered an overlapping strategic priority to move services to primary care, to reduce hospital footfall and improve the carbon footprint and capacity of the care pathway. The GP participant gave several examples in which their practice had hosted non-ophthalmic specialist outreach services. Primary care optometrists were also open to this form of decentralisation but held reservations due to their experience with several prior stalled or failed initiatives to do so for other clinical services.

4.6.6 Wider system

At the time of data collection, it was already common across all stakeholder groups for negative associations with other forms of AI (and often robotics) to be volunteered from public media and science fiction sources. Even following further explanation of the technology on which this study focused, these associations often resurfaced within discussions and appeared likely to colour peoples' perceptions of AI-enabled care.

> *"I think rightly so we should be frightened of it [AI]. You know, we've all watched films like 'Number 9' and all those kinds of really spooky... And I think we do need to be very respectful of that. And of people's fears of it as well. You know, look at the response to vaccination and the amount of distrust that it uncovered."* [Manager]

This senior leader within the hospital emphasised that these perceptions should be accommodated in any AI implementation plans as they have serious consequences, even if their evidence base is unclear. Patient and charity participants offered some suggestions on how patients may be enabled to shape or build trust in any potential clinical AI interventions. Charity professionals self-identified their role as surfacing the patient voice and advocating for it across various stakeholder groups who drive health service innovation. It also appeared that they had stronger connections to national regulatory and commissioning gatekeepers than many hospital clinicians and managers.

> *"We're quite involved in policy and some regulatory discussions as well. So we do work quite closely with NICE and we are engaging more with different parts of the NHS and some of the strategy development around these pathways being designed. Having a clearer sight of what's on the horizon and what's being developed and how it works and what it means for our patients*" [Charity professional]

This perspective from the charity's senior leadership was echoed by a regional representative who wanted to find ways to strengthen this bridging and advocacy role at a local level. This came with the acknowledged shortcoming in engaging with "*anyone who's not a relatively wealthy, middle class, white lady"*, addressing which was a firm priority for the charity. Community optometrist participants also raised concerns that different groups of patients may find it more challenging to access nAMD services. These groups were mainly people for whom repeated journeys to hospital were particularly challenging (more comorbidities, social isolation, rurality or financial pressures). For that reason, greater decentralisation of nAMD care, whether facilitated by AI implementation or not, was seen as a likely means to improve service equity.

> *"…in an affluent area you may well get three hubs within a small [area] fully staffed, and somewhere where perhaps there are not as many staff. You may well have the set up, you may well have the go ahead, but you don't have the staff members to roll it out."* – [GP]

This participant highlighted that the backdrop of staff shortages still threatened service equity in a decentralised model, as clinicians may choose to take jobs closer to the affluent areas in which they tend to live. Several participants highlighted that improving healthcare equity was another key priority at national and regional levels for the NHS and so suggesting or evidencing its delivery with AI-enabled nAMD care could be influential in gaining support from leadership.

4.6.7 Embedding and adaption over time

Most participants highlighted a sustained need to improve NHS capacity for nAMD care. It was expected this need would increase over time. Similarly with AI, industry and NHS professional participants all felt that the context in which they were working was currently highly supportive of AI-enabled healthcare innovation.

> *"I would say that ministerially and in other aspects, there is a need to get something [an AI-enabled care pathway] out because there are short-term objectives of political will."* [Policy maker]

This national policy lead explained the political drive to achieve some widely recognised examples of successful clinical AI implementation. It was also emphasised that these examples were likely to be achieved in clinical areas with pre-existing high digital maturity or need for change. The scale of the short-term impact was therefore thought to be limited as a majority of clinical settings had more fundamental digital issues (e.g. electronic healthcare record implementation) before clinical AI implementation would become a realistic priority. Alongside other perspectives from industry and commissioners, these questions over the scalability of AI benefit across the NHS suggested that the future of the specific strategic emphasis on AI as a solution type may be relatively fragile and that its sustainment should not be assumed. Uncertainty about future directions could also influence the scale and ease of investment made in clinical AI development. As clinical AI begins to be implemented, a practical understanding of the challenges of delivering commercial and clinical benefit are becoming clearer.

> *We often as consultants feel a bit like bereavement counsellors. Because we have very excited people come in, with some great research, with great potential, but the reality is that this [AI medical devices] is a very heavily regulated space, time consuming and expensive"* [Regulatory professional]

This regulatory consultant also imparted that most clinical AI product development is performed by independent small young companies who most commonly use external investment to absorb the associated risk and expense. This was mirrored by industry participants, all of whom were employed by large companies invested in clinical AI strategies. These strategies seemed to be oriented more toward system solutions (e.g. cloud platforms and interoperable electronic healthcare records) capable of aggregating the success of a range of AI products.

> *"You need to work with the existing system that you have, which is massive, convoluted, complicated, and is regulated both by government and other bodies. And introducing changes to such an organism is not easy at all. So, we will see it [AI implementation] as an iterative process in my mind."* [Industry professional]

This international strategic lead for a MedTech company shared their appreciation for the complexity of implementing AI and suggested that better informed strategies would become possible as the fate of the current wave of interest in AI emerges over time.

4.7 Discussion

Besides the data it elicited, this chapter served to validate the adaptions to the NASSS framework made in the qualitative evidence synthesis of chapter 3 and the stakeholder groups which were abstracted. The data presented are supportive of these outputs from chapter 3, as insights from different stakeholder groups recruited were unique in at least some regards and no new framework subthemes were deemed necessary through the analysis process. Furthermore, the interdependence of factors that could influence the implementation of AI-enabled nAMD treatment monitoring was well illustrated by the recurrence of themes across several NASSS domains. This is well illustrated by the potential shift of nAMD care to primary care, which related to most if not all the seven domains.

There appear to be many factors which are likely to positively influence efforts to implement AI-enabled nAMD treatment monitoring, including:

- trust that AI-enabled care will not reduce visual outcomes
- strategic priorities of decision makers (capacity increases, decentralisation, equity, sustainability etc,)
- ophthalmologists' and patients' drive to reduce the volume of consultation episodes

The qualitative interview findings also revealed challenges, including:

- regulatory hurdles that products and manufacturers are required to clear and the limited awareness of them from most stakeholders
- current digital infrastructure and expertise of potential NHS adopter organisations
- cultural scepticism for AI in general terms

This rich dataset helps to understand what factors could influence the implementation of AI-enabled nAMD care and why. In particular it is clear that much of the value proposition depends upon the performance of AI in relatively autonomous monitoring of nAMD treatment. If this evidence cannot be produced it seems that research investments would be best placed in further developing the technology rather than the intervention in which it should sit. Gaining practical value from insights on what could influence implementation and why, to inform how AI-enabled nAMD care should be implemented, will then become an important next step.

4.7.1 Comparison with prior work

The central influence of consultant ophthalmologists over other stakeholders was clear throughout the interviews, even though they hold relatively little direct decision-making power regarding adoption. As such it was interesting that the two consultant ophthalmologists interviewed (1 retina specialist and 1 commissioning lead) framed the AI as a means for consultants to extend the scope and consistency of a consultant's approach to practice, rather than disrupting it. This logic of extension, as opposed to the more commonly perceived logic of disruption, has been clearly described in a Norwegian primary qualitative study of a rule-based clinical AI system for oncology.[24] Here, the clinical participants embraced the tool because they felt it deployed the same decision-making mechanisms that they did in their personal practice and welcomed the opportunity to extend their scope of application. This contrasts with US histopathologists' perceptions of an AI-enabled diagnostic aid with relatively low transparency or explainability. Whilst the clinicians were impressed with the performance of the model, they found it hard to trust because they couldn't relate to the process by which outputs were generated.[25]

There were also concerns raised about the readiness of NHS digital infrastructure, but also the clinical and technical NHS workforce to enable AI implementation. The concerns have been long-anticipated and have been the focus of high-profile policy documents commissioned by the UK government.[26-28] The most recent of these focuses on 5 different roles or 'archetypes' that NHS staff will fall into in their interactions with clinical AI and 3 categories of competencies, or learning outcomes that they may require; general digital health competencies, foundational understanding of AI technologies and competencies specific to a particular clinical AI product.[27] In the NHS staff involved and discussed in the present chapter it seems that both general AI understanding (e.g. how to evaluate an AI product) and specific competencies relating to the specific AIaMD embedded

within an AI-enabled macula service (e.g. understanding what applications are within its regulated intended use) will be important elements of any intervention.

Many stakeholders expressed a need for clarity on who would be accountable for AI-enabled clinical decisions. The clinical AI literature mainly focuses on product development and validation, leaving this relatively unexplored. There are very occasional academic or policy publications around procurement practice and very little about long term monitoring or lifecycle management of clinical AI.[29] This has recently improved with the publication of a British Standard for healthcare providers evaluating AI products and an 8-step process from problem identification to clinical AI decommissioning drawing on the organically developing practices of US academic medical centres.[30, 31] There is very little precedent for clinical AI adoption in the UK and therefore how these processes will play out in practice remain unclear. One leading centre, University Hospitals Birmingham NHS Foundation Trust, uses the medical algorithmic audit (MAA) as a framework to adopt a systematic approach to clinical AI monitoring to inform decisions around its use.[32] This same framework will be used to examine the processes that providers may need to adopt if they are to safely implement AI-enabled nAMD care in chapter 7.

Tailored evaluations of the nAMD monitoring tool, were clearly valued by potential adopters. Value was ascribed to a close relationship between the contexts in which the evaluation took place and which the adopters routinely experience. For example, one patient participant even said they would want to see AI outputs for their own prior consultations before they would trust it. This preference aligns with literature elsewhere, including the aforementioned 8-step process where the 5th step, immediately prior to clinical integration, involves local non-interventional evaluation.[30] This is mirrored in several different studies which describe months of 'silent trials' for AI products before they 'go live'.[14] These silent trials are not simply trust-building exercises. They have practical value in facilitating the identification of real-world problems. These problems could cause patient harm if the 'go live' phase had started without their identification and mitigation.[14] Similar small-scale or simulated workflows can also help to check assumptions. For instance, many stakeholders assumed efficiency to be part of the value proposition of AI-enabled care, but this assumption is yet to be evidenced. This practice of 'silent trials', or a similar approach, is certainly desirable in the UK, but may be challenged to some extent by cultural, legal and technical barriers between AI vendors and healthcare providers.

4.7.2 Limitations

A major limitation of this study was its failure to recruit any patients or carers from an ethnic group besides 'White British'. Two patients of Chinese and Indian ethnic backgrounds respectively were consented to participate (in one case with the support of a Cantonese interpreter) but both declined to arrange an interview on subsequent follow-up. Upon searching the available records at the host site, this represented the only current nAMD patient with an interpreter booked for their appointments and 2 of the 8 patients who had been treated for nAMD in recent years with an ethnicity other than 'White British' or 'Not Specified' on the electronic medical record (EMR). Following discussion with the reference and study advisory groups, a mitigation strategy was deployed to recruit charity professionals and community optometrists serving more ethnically diverse localities, hoping that they could advocate for any patient perspectives which may be influenced by ethnicity.

nAMD affects older patients and the study recruited from the Newcastle and Northumbria wards which returned recent census data suggesting the populations aged over 65 suggesting are 95.3% and 99.3% White British respectively.[33] This skewed representation also has a biological drive as nAMD is a quarter and third as prevalent in Black and Asian ethnic groups respectively.[34] These factors mean that the low ethnic diversity of nAMD patients is not in itself evidence of inequities in access and/or uptake of nAMD services by ethnicity. However, this finding motivated, and informed the design of, a local quality improvement project to improve the completeness of ethnicity recording on the EMR (see appendix).

Another key limitation to the relevance of this work is its focus on clinician, manager, patient and carer stakeholders at a single macula service. There was some hint at the high degree of variability in service design between different macula services from participants who drew on regional, national or even international experiences. This was an intentional choice to facilitate a deep exploration of a single implementation context, rather than an inevitably more superficial yet generalisable exploration across multiple UK centres. This choice to focus on a tangible implementation site also aimed to avoid highly abstract and unactionable findings, which can result from qualitative explorations of hypothetical innovations such as the one under study.[12] To mitigate against the resulting limit to generalisability, NASSS was also used to abstract the study's findings to make them more relevant to implementation efforts outside of Newcastle upon Tyne Hospitals NHS Foundation Trust (NuTH).[12]

As demonstrated by chapter 3, this primary qualitative study represents the most comprehensive sampling of stakeholder groups to AI implementation in any single study.[13] Despite that, there were stakeholders who proved to be infeasible to recruit that may well have added to the study's findings. Individuals who were pursued for participation but, a suitable contact couldn't be identified, did not reply or declined included; senior IT staff, senior clinical informatics staff, an Integrated Care Board (ICB) digital health representative and a representative of the Care Quality Commission. Even within stakeholder groups some limitations will be derived from a sub-optimal diversity in representation, e.g. both carer participants were children of patients. Individuals who were successfully recruited helped mitigate against these limitations by contributing their partly overlapping perspectives, but the study findings are likely to fail to draw on at least some relevant perspectives or insights.

4.7.3 Future directions

4.7.3.1 Within scope of this thesis

The present chapter has identified *what* can be expected to influence the implementation of AI-enabled nAMD treatment monitoring and *why*. The analysis process has allowed these implementation determinants to be distilled from nearly 29 hours of interviews from 36 varied participants. However, this list of determinants offers little assurance that candidate AI technologies can meet minimum requirements for stakeholder acceptance. Foremost among these is a requirement that introducing AI should not compromise the visual outcomes that patients can currently expect. Testing if a potential AIaMD could meet this requirement with real-world clinical data will form the aim of chapter 5.

The determinants surfaced within this chapter offer little practical detail of *how* AI-enabled nAMD treatment monitoring should be implemented. Designing an AI-enabled intervention which aligns with the determinants discussed in the present chapter would help to answer this question of *how* the technology should be implemented. A TMF with a practical focus on healthcare interventions will be used in chapter 6 to propose an evidence-based intervention to carry the AIaMD (to be validated in chapter 5) into clinical practice.[35]

4.7.3.2 Outside scope of this thesis

Whilst nAMD is the major single disease contributing to demand in NHS macula services, there are other retinal diseases which make major contributions and are likely to be amenable. The factors that influence implementation relating to the condition and adopters are at least partly distinct however, and so additional qualitative research would hold value. These diseases include diabetic macula oedema, retinal vein occlusion, myopic macular degeneration and geographic atrophy in AMD. Designing an AI-enabled intervention to support the treatment of all of these diseases (or variant interventions for each) would increase the potential impact of the technology under study.

To support the generalisability of this chapter's findings across the NHS, it would be beneficial to vary the context from which participants are recruited in further work. This could prioritise macula service representation from across the four UK nations, independent sector services and small rural and large urban services. To maximise the scope of such work it may be valuable to conduct interviews or focus groups with sites and services different to NuTH and then test the range of determinants in a national survey. Given the widespread concerns about differential performance across ethnic groups, it would also be helpful to collect data from services that serve a more ethnically diverse population.[36]

4.8 Conclusions

There are many determinants of implementation success for AI-enabled nAMD treatment monitoring. The potential value proposition for AI in this context extends far beyond the motivating opportunity to improve capacity in nAMD services. It includes improvements to patient and clinician experience, healthcare equity, carbon footprint, care quality and care consistency. There are also many different stakeholder groups who could influence implementation, but consultant retina specialists are of central importance, with many other stakeholders looking to them to help form their own opinion. All stakeholder groups see the preservation of visual outcomes in nAMD as a minimum requirement for AI technologies to be implemented. The absence of such evidence currently represents a barrier to implementation. Whilst this chapter's findings offer insights on *what* could influence AI-enabled nAMD treatment monitoring and *why*, the details of *how* implementation should be conducted remains unclear.

4.9 Appendix

4.9.1 Example topic guide (patient)

Current pathway

1. What parts of the current service do you like most?
2. What changes would you like to see in the current service?
3. In the current service you meet lots of team members for different parts of your care. Which bits of face-to-face or written contact do you find most valuable?

4. Have you had any ‘injection only’ or virtual clinic visits, where a consultation isn’t part of the visit? If so, how do you find them?

Initial impressions of clinical AI

1. As you know, I’m interested in the idea of technology in healthcare. What are your first thoughts and feelings when I mention artificial intelligence in healthcare?
2. When I’m talking about AI here, I mean a type of technology designed to take in some information and make a decision about it on its own. Specifically, I’m talking about a kind of AI we have, which can look at your eye photos on its own and see when your eye next needs an injection without help from a nurse or doctor. What use could you imagine for AI like that?
3. What kind of down-sides or difficulties do you think there might be in using that kind of AI in the clinic?

Pathway placement

1. Who do you think the best person to be responsible for the AI would be?
2. What kind of interactions would you like to have with doctors if artificial intelligence is brought in?
3. Where would you like to have your eye photos taken?
4. Where would you like to have your injections given?

Relationships with the tool and others

1. Some people might feel a bit uncomfortable about letting AI take some of the responsibility for their treatment planning. What kind of things might help you trust AI like this?
2. With this kind of AI, you might be able to see how and why it makes its treatment decisions for you. How would you feel about that?
3. How might it change the way friends and family support you in managing your eye disease?
4. Who would you want to be able to access the AI’s decision making?
5. How do you think bringing this AI into the service would affect your relationships with different members of the care team?

Closing

1. Thanks very much for so many helpful insights. Is there anything else we haven’t talked about that seems important about using AI in macular degeneration clinics?
2. We’re planning on talking to hospital doctors and nurses, opticians and managers but do you think there are other people’s perspectives we should be hearing?

4.9.2 Example reflective journal

“This is just a quick reflective diary after the interview with patient three, which felt like it went well. I think I did a better job of letting them speak… certainly consciously. There were a few times where I wanted to jump in, but I didn't. And actually, a lot of what I wanted to be said was said, in a way. So that's good. And just coming back to the kind of demographic,

I think it's, it's well highlighted by the fact that two out of my three patient members I've recruited so far play croquet. I think that probably points to a fair bit of bias towards wealthier patients that's emerging in the sampling. I don't want that, but in a way it's helpful because they're quite empowered patients and good advocates for their own perspectives and people around them as well. So it's not all bad, but it's probably not representative of the full population. So that's something we need to address. I don't think there's any other major issues. I guess that patient three had obviously done quite a lot of research into AI himself, both to actual kind of formal research and discussion with family members. So that may well have limited the representativeness of his perspectives, but there were some interesting points in that as it kind of flags this idea that the general public needs to come around to the idea, read more about it themselves and understand the issues themselves. So that was interesting. And the other thing that came up before the recording started was that we discussed his appointments at the beginning, he was talking about how he has just been chasing his appointment so that was a good example of how service capacity is key. He can't get the appointment that he needs right now. So it's a nice illustration."

TEMS

TECHNOLOGY ENHANCED MACULAR SERVICES

Reference Group Handbook

Study Lead:

Jeff Hogg

NIHR doctoral fellow

Jeffry.Hogg@newcastle.ac.uk

Newcastle University

NHS

The Newcastle upon Tyne Hospitals

NHS Foundation Trust

NIHR | National Institute for Health Research

Background

Using UK taxpayer money, the National Institute for Health and Care Research are funding a 3-year research project called 'A mixed methods validation of Technology Enhanced Macula services'. Eye care is the largest contributor to NHS hospital clinic appointments and in turn the largest contributor to these appointments is age-related macular degeneration (AMD). The amount of people needing treatment for this disease is rising and the resources that the NHS have are already failing to meet these needs. If a solution is not found then treatment will be delayed and people will lose vision that could have been saved.

This research project aims to see if a type of artificial intelligence, which has been developed by a separate team, could solve this imbalance of supply and capacity in services for people with the type of AMD we can treat, exudative AMD. We aim to do this in 3 steps. Firstly, we will summarise all the published evidence on how people involved in all sorts of healthcare services felt when they used technologies that support clinical decisions. Second, we will interview patients and healthcare professionals involved in AMD clinics. Finally, we will use photos from past AMD clinics to see what treatment artificial intelligence would have suggested and compare that to the decisions made by NHS staff.

Purpose of the Reference Group

This research is funded by the public and hopes to make meaningful improvements to the services we can all access. The purpose of the Reference Group is to support this goal, by preventing the project being driven purely from the perspective of researchers and clinicians. The intention is to discuss upcoming steps of the project and ensure that plans and interpretation align with public values as much as possible. Ideas generated in these discussions will feed directly into the management of the project, either directly or with additional input from the Study Advisory Group.

Objectives

To ask research questions that will help to develop AMD services that are both convenient and effective for members of the public.

To help to interpret the meaning of interviews and research data without academic or professional bias.

To inform ways of communicating research findings to the public in an accessible and engaging way.

Study oversight

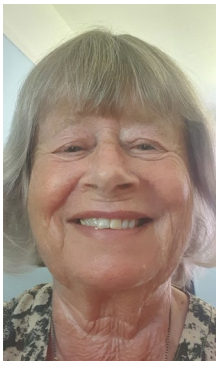
Before work started a detailed proposal for the TEMS research study was put together with input from members of the public and topic experts and was approved by the funder, the National Institute for Health and Care Research. During the three years of proposed research, guidance will be provided by two separate groups to maintain the public value and the scientific quality of the project. The first of these is the Reference Group, consisting of four members of the public experienced in supporting health and care research in other contexts. The second is the Study Advisory Group, where the membership have varied clinical, academic and lived experience related to this specific research project. Each group will aim to meet twice yearly to discuss the project, with flexible communication between these points.

Study lead

Jeff Hogg 	I am a junior doctor specialising in ophthalmology who is taking time out from clinical practice to study for a PhD at Newcastle University's Population Health Science Institute. I have lived and worked in Newcastle for more than ten years and gained experience in observational and diagnostic research. Supported by the team listed in this document and others, I formed the proposal for this study in response to the rising volume of appointments associated with the macula service for the NHS, clinicians and patients. Whilst artificial intelligence that could streamline these services for all stakeholders is relatively mature, my passion lies in understanding how to bring it into practice in a way that works for everyone.

Reference Group members

Rashmi Kumar	I am now a fulltime Carer for an elderly mother suffering from Long term multiple illness. This experience has helped me to better understand and appreciate some of the health, psychological and social challenges patients (and their families) face in their lives every day. Crucially, it has helped me understand how better support could significantly improve their health and wellbeing. I am from BAME culture and a Trustee of large Patients Participation Group (PPGs) Network in South London which has very diverse and extensive BAME communities, with many experiencing significant health and social care

	deprivation. I am actively engaging with many Primary Care Networks, GP Federations and the South East London CCG. I am also a member of Cicely Saunders Institute of Palliative Care, Policy and Rehabilitation, which focuses on research on improving health and social care services for patients, their families and support networks, on management of Palliative Care and EoL support services. Coming across this study on the Technology Enhanced Macula services, particularly on the potential determination of the Age-related Macular Degenerations (AMD), I feel this could be a useful study to explore and develop potential improvements on health and wellbeing of some communities which may not be receiving optimum level of care and support they deserve.
Christine Sinnett 	I have now retired after working in administration since the age of 16. I completed a BA in Sociology followed by an MA in Contemporary Sociology at Durham University, geared towards research, which was awarded in 2002. I am a member in VOICE Newcastle and have participated in many research projects such as the current TEMS Project which is proving to be extremely interesting. My interest in this research was driven by family occurrences of MD. My uncle with wet AMD had to travel a long way from Brampton for some treatment. My cousin currently developed a hole in her macular which thankfully appears to have receded, but she has had a lot of imaging at the RVI. Anything that can be done to alleviate the logistical problems of having to travel and wait for treatment can only be a good thing.
Rosemary Nicholls	I was delighted to hear that my application to join the Reference Group had been successful, as I have found my experience as a member of the North-East Research Design Services Consumer Panel very rewarding and would like to make this further contribution to health research. In my younger days, I took several courses in Social Research Methods and have been able to use what I learned then, as a lay person considering project summaries. Although I enjoy very good health, with no experience of eye disease, I know others who are regularly checked for glaucoma

	and macular degeneration. I'm very conscious of the important role sight plays in ensuring my independence and keen to support this promising research project.
Angela Quilley 	I had a career in education, at all levels, enabling me to specialise in several areas including developmental psychology, specific learning difficulties and community education. I have trained nursery nurses to work in hospitals, schools, nurseries and care settings. Most recently I taught children with a diversity of special needs and disabilities. I worked on the compilation of pupils' Education Health and Care plans and psychometric testing to enable students' equality of access to the curriculum and public examinations. This person centred approach enabled me to realise the significance and importance of Public Patient Participation in HealthCare and Health research, which I have embraced for several years. I am part of a multidisciplinary co-production called "Hearing Birdsong" which has developed a user friendly, patient centred way to access help with hearing loss. As a member of the Public Partners Advisory Panel for the NIHR Applied Research Collaboration North West London, I am able to engage in a variety of aspects of research.

Study advisory group

Fiona Beyer 	Fiona leads the information programme at Newcastle University's Population Health Science Institute. She has worked on systematic reviews for over ten years both as an information specialist and a systematic reviewer.
Katie Brittain	Katie is professor of applied health research and ageing at Newcastle University's Population Health Science Institute. Her recent work has focused around how aspects of the physical, social and technological environment pose challenges and opportunities for older people and their wider community.

	
Pearse Keane 	Pearse is a consultant ophthalmologist at Moorfields' Eye Hospital in London where he specialises in the medical treatment of retinal disease. He is also professor of artificial medical intelligence at University College London's Institute of Ophthalmology.
Trevor Lunn	Trevor is a retired general practitioner who worked and continues to live in the North-East. He has several years of experience of the local macula service as a patient and is a member of the Macular Society.
Janet Lunn	Janet is a retired nurse who worked and continues to live in the North-East. Through her husband Trevor she has substantial experience of the local macula service and supporting someone with macular degeneration.
Gregory Maniatopoulos	Greg is an assistant professor in healthcare innovation at Northumbria University. His research interests lie primarily in the broad area of health systems, implementation and change, in particular exploring how organisational, technological and policy factors shape processes of appropriation of innovations in healthcare practice.
James Talks 	James is a consultant ophthalmologist at Newcastle upon Tyne Hospitals where he leads the medical retina service and macula service.
Dawn Teare	Dawn is professor of biostatistics at Newcastle University's Population Health Science Institute. Nationally she co-leads the NIHR statistics group and is experienced in biomedical research, clinical trials and research integrity.

4.9.4 Exemplar codebook summaries

4.9.4.1 Type or format of care need

Patients, clinicians and commissioners perceived a very high personal and economic burden from nAMD treatment. This comes from both the frequency of injections and the serious consequence of irreversible sight loss if that frequent need is not met. Coupled with the high prevalence of the condition, nAMD treatment was a high strategic priority for commissioners and hospital managers. There is also a sense that treatment would be improved if the act of scheduling and administering the treatment did not have to displace more patient-centred social, psychological and broader clinical considerations. At present, most of clinicians' attention is diverted toward the relatively simplistic decision-making around when treatment is required for patients with established diagnoses:

"…the two decisions that you need to make are, do they need an injection, and when do they need the next one?... You just look at the picture and say it's dry, and they've had 10 weeks, let's try 12 weeks… It's as simple as that. So, you don't need a person." [HCP4]

Here an ophthalmology trainee shares their frustration over the amount their time spent on producing these decisions around treatment timing, later stating "my daughter can tell you that". Similarly, patients, clinicians and mangers felt that the delivery of the injections themselves is an inefficient use of ophthalmologists time. Both patients and HCPs expressed a preference to prioritise this time for consultations around diagnosis, screening for ocular co-morbidities and changes to the management plan (e.g. cessation of injections).

4.9.4.2 Tools redefine staff roles

Most participants felt that AI-enabled nAMD treatment monitoring would change both the nature of work required and the staff groups best-suited to the work.

"If you are deskilling them in a role that's no longer needed, well, that's not a problem. We no longer need the person who looks after leeches. That's not a big deal." [HCP 11]

This GP hints at the extreme end of this spectrum with potential redundancy for eye specialists to be involved in nAMD treatment monitoring. Some participants felt an unfair professional threat for themselves or others, given the personal investments HCPs had made to achieve their competencies. Patients felt these shifting roles could assign greater priority to discussion and empathy from their clinicians. Most HCPs expected the potential for role expansion through AI-enabled care to increase the value they could contribute to patient care. Commissioners voiced things more pragmatically but seemed to welcome the opportunity to reduce staffing requirements for care provision. Most participants caveated their contributions, recognising that the exact implications of AI adoption would depend upon the detail of the use case.

4.9.5 Quality improvement project

Ethnicity recording rates in the Royal Victoria Infirmary Macula Service (reported and led by Dr Samy El Omda, supervised by Dr Jeffry Hogg)

Introduction and background:

COVID-19 shone a harsh light on some of the health and wider inequalities that persist in our society. NHS England responded to this with an initial 8 urgent actions for tackling health inequalities, which was later refined to 5 key priority areas. Priority 3 states: "Ensure

datasets are complete and timely – Systems are asked to continue to improve the collection and recording of ethnicity data across primary care, outpatients, A&E, mental health, community services, and specialised commissioning".

This is further supported by the NHS 10-year long term plan in which strategic plan 2 is "Preventing illness and tackling health inequalities.

Aims and objectives:

We aim to see what percentage of patients at our Age Related Macular Degeneration (AMD) clinics have their ethnicities recorded and how we can improve this data if we are not meeting the standards.

For Audit:

Title of standard document: 2021/22 priorities and operational planning guidance: Implementation guidance 25 March 2021

Specific defined standards to be measured: Percentage of patients with ethnicity recorded.

Methods:

Data was collected via Medisoft from AMD + injection clinics up to December 2022. Patients MRN and ethnicity were extracted. We then analysed the percentage of each ethnicity category recorded.

Results:

90.4% of all patients had their ethnicity recorded with 9.56% currently being not stated.

Conclusion:

Currently although we are achieving good ethnicity recording rates, especially in comparison to the trust average of 80% of all patients having their ethnicity recorded, there is still room to improve. This is of particular importance given Newcastle's ethnicity breakdown in the 65+ age group (where the majority of patients with AMD fall into) as with 97% of this population being identifying as "White" it is important we have very accurate ethnicity data if we are to do any statistically significant analysis on how well we are reaching the other 3% of Newcastle's Population.

Recommendations for the future:

We need to identify the current barriers to 100% ethnicity recording.

Action taken to disseminate QI/audit findings:

Presented findings at the Receptionist meetings – Identified issues they had with recording ethnicity. 1) Patients not understanding why ethnicity is being asked and hence being apprehensive around answering this question. 2) Clinics being very busy and there not being time for this without delaying the clinic.

Actions to address areas of development:

Following this advice, we discussed with one of the ethnicity leads at the RVI and created an information sheet with a questionnaire that explains the importance of ethnicity recording and allows the patient to tick what ethnicity they identify with and hand this back to the desk to be recorded at a later time.

Intervention started 16th October 2023

Re-audited 31st October 2023

As our macula clinics run every day, we re-audited this after 2 weeks for preliminary results and to see if the reception team identified any issues. Although we were informed that most patients in these clinics had their ethnicity recorded already, the intervention worked successfully achieving 100% ethnicity recording tin these two weeks.

4.10 References

1. Beede, E., et al., A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020, Association for Computing Machinery: Honolulu, HI, USA. p. 1–12.

2. Held, L.A., L. et al, Determinants of the implementation of an artificial intelligence-supported device for the screening of diabetic retinopathy in primary care - a qualitative study. Health Informatics J, 2022. 28(3): p. 14604582221112816.

3. Robinson, E.L., et al., Artificial intelligence-integrated approaches in ophthalmology: A qualitative pilot study of provider understanding and adoption of AI. Investigative Ophthalmology & Visual Science, 2022. 63(7): p. 729 – F0457-729 – F0457.

4. Gunasekeran, D.V., et al., Acceptance and Perception of Artificial Intelligence Usability in Eye Care (APPRAISE) for Ophthalmologists: A Multinational Perspective. Front Med (Lausanne), 2022. 9: p. 875242.

5. Giocanti-Aurégan, A., et al., Drivers of and Barriers to Adherence to Neovascular Age-Related Macular Degeneration and Diabetic Macular Edema Treatment Management Plans: A Multi-National Qualitative Study. Patient Prefer Adherence, 2022. 16: p. 587-604.

6. Talks, S.J., et al., The Patient Voice in Neovascular Age-Related Macular Degeneration: Findings from a Qualitative Study. Ophthalmol Ther, 2023. 12(1): p. 561-575.

7. Sii, S., et al., Exploring factors predicting changes in patients' expectations and psychosocial issues during the course of treatment with intravitreal injections for wet age-related macular degeneration. Eye (Lond), 2018. 32(4): p. 673-678.

8. Calles-Monar, P.S., et al., Modifiable Determinants of Satisfaction with Intravitreal Treatment in Patients with Neovascular Age-Related Macular Degeneration. Drugs Aging, 2022. 39(5): p. 355-366.

9. Vukicevic, M., et al., Caregiver perceptions about the impact of caring for patients with wet age-related macular degeneration. Eye (Lond), 2016. 30(3): p. 413-21.

10. Hull, L., et al., Designing high-quality implementation research: development, application, feasibility and preliminary evaluation of the implementation science research development (ImpRes) tool and guide. Implementation Science, 2019. 14(1): p. 80.

11. Birken, S.A., et al., T-CaST: an implementation theory comparison and selection tool. Implementation Science, 2018. 13(1): p. 143.

12. Hogg, H.D.J., et al., Unlocking the potential of qualitative research for the implementation of artificial intelligence-enabled healthcare. Journal of Medical Artificial Intelligence, 2023. 6.

13. Hogg, H.D.J., et al., Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence. J Med Internet Res, 2023. 25: p. e39742.

14. Kwong, J.C.C., et al., The silent trial - the bridge between bench-to-bedside clinical AI applications. Front Digit Health, 2022. 4: p. 929508.

15. Curran, G.M., et al., Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care, 2012. 50(3): p. 217-26.

16. Greenhalgh, T., et al., Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res, 2017. 19(11): p. e367.

17. Nilsen, P., Making sense of implementation theories, models and frameworks. Implementation Science, 2015. 10(1): p. 53.

18. Brinkmann, S., Unstructure and Semi-Structured Interviewing, in The Oxford Handbook of Qualitative Research, L. P, Editor. 2014, Oxford University Press: Oxford, UK. p. 277-279.

19. Kamberelis, G, and Dimitriadis G, Focus Group Research: Retrospect and Prospect, in The Oxford Handbook of Qualitative Research. 2014, Oxfor University Press: Oxford, UK. p. 315-340.

20. Manzano, A., The craft of interviewing in realist evaluation. Evaluation, 2016. 22(3): p. 342-360.

21. Ortlipp, M., Keeping and Using Reflective Journals in the Qualitative Research Process. The Qualitative Report, 2008. 13(4): p. 695-705.

22. Saldana, J., Coding and Analysis Strategies, in The Oxford Handbook of Qualitative Research, P. Leavy, Editor. 2014, Oxford University Press: New York. p. 581 - 605.

23. American National Standards Institute, Definitions/Characteristics Of Artificial Intelligence In Health Care - ANSI/CTA-2089.1-2020. 2020.

24. Torenholt, R. and H. Langstrup, Between a logic of disruption and a logic of continuation: Negotiating the legitimacy of algorithms used in automated clinical decision-making. Health (London), 2023. 27(1): p. 41-59.

25. Cai, C.J., et al., "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. Proc. ACM Hum.-Comput. Interact., 2019. 3(CSCW): p. Article 104.

26. Goldacre, B., Better, broader, safer: using health data for research and analysis. 2022.

27. NHS AI Lab, NHS Transformation Directorate and NHS Health Education England, Developing healthcare workers' confidence in artificial intelligence (AI) (Part 2). 2022.

28. Topol, E.J., High-performance medicine: the convergence of human and artificial intelligence. Nat Med, 2019. 25(1): p. 44-56.

29. NHS X, A buyer's guide to AI in health and care, T. Directorate, Editor. 2020.

30. Kim, J.Y., et al., Organizational Governance of Emerging Technologies: AI Adoption in Healthcare, in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023, Association for Computing Machinery: Chicago, IL, USA. p. 1396–1417.

31. Sujan, M., et al., Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. BMJ Health Care Inform, 2023. 30(1).

32. Liu, X., et al., The medical algorithmic audit. The Lancet Digital Health, 2022. 4(5): p. e384-e397.

33. Offie of National Statistics, Population estimates. 2021.

34. Zhou, M., et al., Geographic distributions of age-related macular degeneration incidence: a systematic review and meta-analysis. Br J Ophthalmol, 2021. 105(10): p. 1427-1434.

35. Hogg, H.D.J., et al., Evaluating the translation of implementation science to clinical artificial intelligence: a bibliometric study of qualitative research. Front Health Serv, 2023. 3: p. 1161822.

36. Obermeyer, Z., et al., Dissecting racial bias in an algorithm used to manage the health of populations. Science, 2019. 366(6464): p. 447-453.

Chapter 5: A retrospective non-inferiority study of an AI-enabled tool for nAMD treatment monitoring versus consultant-led-care

Problem: There are no examples of clinical AI technology being used to make treatment plans for nAMD that are non-inferior to current clinical practice. Chapter 3 identified this as a minimum requirement for many stakeholders and without establishing whether it has been satisfied it is unclear if further investment to improve the technology and design or evaluate the intervention in which it will sit are warranted.

Objectives: To test the non-inferiority of a potential AI medical device for nAMD treatment monitoring against consultant-led care at NuTH.

Methods: Single-centre retrospective imaging and clinical data were collected from 262 nAMD clinic visits at NuTH, including judgements of nAMD disease stability or activity made in real-world consultant-led-care. Outputs from an AI-enabled retinal segmentation tool were processed by a rule-based decision tree to independently analyse imaging data to report nAMD stability or activity for each of the 262 clinic visits. Independently, an external reading centre received both clinical and imaging data to generate an enhanced reference standard for each clinic visit. The non-inferiority of the relative negative predictive value (NPV) of AI-enabled reports on disease activity relative to consultant-led-care (CLC) judgements were then tested. A relative clinical non-inferiority margin of 10% was applied.

Findings: Analysis of a pilot dataset enabled the identification of a rule-based decision tree which considered any increase in intraretinal fluid (IRF) as evidence of disease activity. The NPV of the AI-enabled reports were 85.5% (95% CI 77.0%-94.0%) which was higher than that of consultant-led care at 80.8% (95% CI 71.3%-90.4%). This gave a relative NPV (rNPV) of 1.06 (95% CI 0.99 – 1.13), clear of the clinical non-inferiority margin of 0.9. The rNPV was accompanied by a relative positive predictive value (rPPV) of 1.05 (95% CI 0.84 – 1.32) indicating that the AI would not be expected to introduce additional over-treatment for patients. Secondary analyses showed that the clinical non-inferiority of AI-enabled reports appeared robust to comparisons across different professional groups enacting CLC. It also showed that largely through a reduction in false negatives the injection burden for patients and services associated with AI-enabled care was likely to be higher. Retrospective error analysis of the AI-enabled reports identified a number of opportunities to improve performance. With further refinement, a rule set which considered only increases of 10% or more in IRF or SRF to indicate disease activity achieved a rNPV of 1.18 (1.09-1.27) and a rPPV of 1.00 (0.82-1.23).

Conclusions: Reports of disease activity in nAMD based on an example clinical AI technology could offer superior treatment monitoring than CLC without any evidence that over-treatment would increase for patients. To support translation of this technology into practice, an intervention which will optimise the likelihood of successful implementation needs to be designed prior to prospective evaluation.

Relevance to future chapters: The positive result of this evaluation comfortably satisfies the minimum performance standards expressed by stakeholders. This warrants further qualitative analysis to design an intervention to optimise the likelihood of successful implementation of AI-enabled nAMD treatment monitoring.

5.1 Background

The clinical component of chapter 1, described the growing mismatch between demand and capacity in NHS ophthalmology services and the significant role nAMD treatment monitoring plays in contributing to that demand. This demand is also set to greatly expand. Firstly, our population continues to age and therefore the prevalence of nAMD will increase. Secondly, series of injections for geographic atrophy secondary to AMD were approved for use by the FDA on 17th February 2023 and will come to the UK market in 2024 further expanding the proportion of late AMD which can be treated.[1, 2]

Chapter 4 provided evidence that stakeholders in nAMD care would be accepting of additional service capacity to balance demand. Participants expressed uncertainty over the degree of human oversight that should be applied to AI-enabled treatment monitoring but acknowledged that this would strongly influence the ability of any intervention to enhance clinic capacity. These findings mirror discussions elsewhere in the literature that seek to find use-case specific trade-offs between the increasing efficiency saving and risk of clinical harm thought to accompany increasing degrees of autonomy.[3, 4] This means that if AI-enabled nAMD treatment monitoring is to meaningfully enhance clinic capacity, then it must assume at least a moderate level of autonomy and so a high level of independent performance from the AI technology will be required. Evidence of this high level of performance is not just important for decision makers, such as regulators and commissioners, but also for clinical and public end-users who are often sceptical of clinical AI with little or no prior experience to build trust upon.[4] Public and professional conversations around clinical AI, rightly place emphasis on its risk to reinforce or augment pre-existent inequities in healthcare provision too.[5] To build trust with stakeholders and support the implementation of AI-enabled nAMD treatment monitoring, evidence should not just demonstrate high levels of safety for the NHS clinic population as a whole, but demonstrate robust performance in various subgroups. This will help satisfy the need to innovate services in a way that is safe for everyone, not just safe on average.

5.2 Problem

For NHS patients and real-world services to begin to experience benefit from AI-enabled nAMD treatment monitoring, a prospective evaluation producing convincing evidence of its safety and effectiveness is required.[6] The resources and patient risk associated with such a prospective evaluation are not justifiable without supportive evidence from retrospective evaluation. Whilst the analytical validity of various tools to segment retinal OCT is secure, research designed to support the clinical validity of AI-enabled treatment monitoring for nAMD, is yet to be reported.[7, 8] The primary requirement of this evidence is to provide assurances that AI technology with an adequate level of autonomy to enhance clinic capacity, would not compromise patients and other stakeholders' priority to preserve vision.

5.3 Rationale

In a non-interventional study such as the one proposed, it is not possible to directly measure the visual outcomes of AI-enabled nAMD treatment monitoring. In practice, the interval with which anti-VEGF IVIs are given to patients is the main influence on visual outcomes that clinicians can control. Whilst the definitions of disease activity informing anti-VEGF treatment for nAMD are relatively well established, significant variations remain in approaches to service delivery and IVT interval decisions across the NHS (chapter 4).[9] To

maximise the relevance of the present study, a design which holds relevance across these varied services is desirable.

As previously mentioned, avoiding any compromise on visual outcome consistently arises as stakeholders' priority for nAMD care. In the proposed study design, sight loss is threatened by undertreatment and hence the failure to recognise nAMD disease activity. The present study design aims to reflect this priority in its primary outcome, but identifying sight-threatening 'failures' to recognise disease activity is challenged by the retrospective setting of the study. An enhanced reference standard, with broad credibility among stakeholders will be required to evaluate and compare CLC and AI-enabled assessments of disease activity. Moorfields reading centre was selected as such an enhanced reference standard, in light of its track record in ophthalmic research, its reputation among ophthalmology patients and professionals and also regulators acceptance of established reading centres as performance benchmarks.[10] Selecting a level of safety that is good enough is challenged by the novelty of this use case, but assuming that a real-world example of CLC represents acceptable performance a non-inferiority design seems to provide the most relevant evidence.[11]

5.4 Aim

This retrospective study aims to test the non-inferiority of an AI-enabled OCT analysis tool relative to real-world CLC in distinguishing between the presence or absence of nAMD disease activity.

5.5 Methods

5.5.1 Justification of study design and sample size

An enhanced reference standard is required to facilitate comparison between a potential AI-enabled nAMD service and the real-world gold standard, CLC. Moorfields Reading Centre has an international reputation and track record in meeting this need for prior studies and will receive imaging and clinical data for each case in the present study to generate an enhanced reference standard.[12, 13, 16] Reading centre judgements are also accepted by medical device regulators in their process of evaluating products for market, lending pragmatic relevance to this approach. The binary decision under examination for each included visit is whether the data suggest nAMD disease stability or activity. This simplification of the scalar number of weeks between treatments or the three-option decision regarding treatment interval maintenance, extension or reduction is more broadly relevant to real-world practice where clinician and patient preferences about how treatment intervals should be altered on account of disease activity vary.[9] It also lowers the risk of inappropriately labelling a decision made in CLC with subtle influences from patient or clinician preference as 'incorrect'. Whether binary judgements of disease activity (positives) or stability (negatives) from CLC and the AI-enabled decision tool are labelled as true or false was decided by the independent judgement of the Moorfields Reading Centre (Table 10).

Table 10. Template confusion matrix showing the different possible classification of Artificial Intelligence (AI)-enabled reports of disease activity and judgements from consultant-led-care (CLC) for each eligible case.

	Moorfields Reading Centre identifies disease stability	Moorfields Reading Centre identifies disease activity

	(negative)	(positive)
AI identifies disease stability (negative)	AI True negative	AI False negative
AI identifies disease activity (positive)	AI False positive	AI True positive
CLC identifies disease stability (negative)	CLC True negative	CLC False negative
CLC identifies disease activity (positive)	CLC False positive	CLC True positive

Treatment decision data provide good estimates of the design parameters for this novel use case. Considering a different use case, OCTane has demonstrated equivalent or superior retinal diagnostic performance to consultant specialists. Relative to final real-world clinical diagnoses it produced an area under the receiver operator characteristic (AUROC) curve of 99% for retinal diseases including nAMD.[13] This performance was dependent on the same intermediate anatomical segmentation step that will form the basis for the OCTane-based tool for nAMD treatment proposed here. This has supported the feasibility of the current study but does not provide the level of certainty required to perform a robust power calculation, which requires sufficiently accurate estimates of the NPV of both judgements from CLC and AI-enabled reports from paired data.[17] Therefore, the pilot dataset collected was sent for independent processing by Moorfields Reading Centre and the AI-enabled decision tool to supply these estimates. Prior work has established that for binary outcomes such as the one under study, little improvement is seen in estimate precision or bias by increasing the size of the pilot dataset above 100 and so a size between 60 and 100 is recommended.[18] To forecasts the number of cases required, an initial pragmatic estimate of the prevalence of disease activity was made from a focused review of CLC assessments of nAMD activity at 100 NuTH clinic visits under loading or TEX treatment protocols. This found that 79 reported disease stability and would be classed as negative cases, which could contribute to estimating NPV alongside the reference standard. Given that about 79% of eligible visits are classed as negative the randomly sampled pilot dataset required to accrue 100 negative cases was expected to be around 127 cases (100 / 0.79). To try and ensure that the actual pilot dataset contained the necessary 100 negative cases, sampling for the pilot dataset continued until 100 visits with negative CLC judgements had been curated, meaning the pilot dataset size was not fixed until curation was complete. The estimated NPV of AI-enabled reports and judgements from CLC was derived from this pilot dataset and was input to a power calculation. This power calculation first required the categorisation of different types of agreement and disagreement between AI-enabled reports, CLC reports and Moorfields Reading Centre reports (Figure 24).

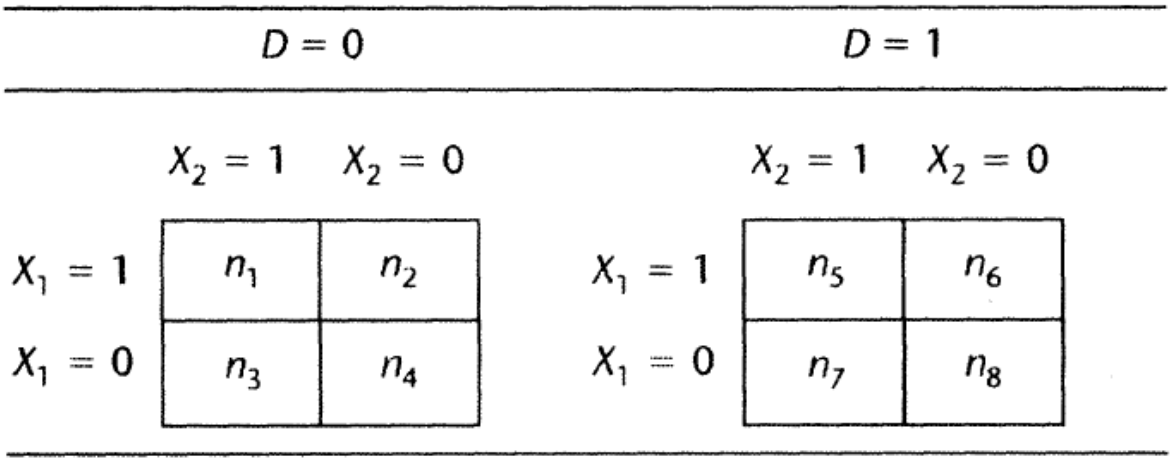


Figure 24. Categorisation (in abstract) of agreements and disagreements between AI-enabled reports (X1), CLC reports (X2) and Moorfields Reading Centre reports (D) of disease activity (1) or stability (0).[17]

The number of cases assigned to each of the 8 potential categories of agreement and disagreement was then converted into a proportion of the whole pilot dataset where p = n/N. With this information, and values for α (accepted risk of type 1 error), β (accepted risk of type 2 error), δ (clinical non-inferiority margin), γ (clinical superiority margin) and the NPV estimate for CLC judgements (X_2), the sample size can then be derived.[17]

Sample size = $[(z_{1-\alpha} + z_{1-\beta})^2 / \ln(\gamma/\delta)]^2 \times 1/[(p_2 + p_4)(p_3 + p_4)]$

$$\times [-2(p_4 + p_8)\gamma NPVX_2^2 + (-p_3 + p_4 - \gamma(p_2 - p_4))NPVX_2 + p_2 + p_3]$$

From this, the number of eligible cases required, in addition to the pilot dataset, was established. These additional cases were sent for processing by Moorfields Reading Centre and the AI-enabled decision tool to test the non-inferiority of AI-enabled report NPV relative to the NPV of judgements from CLC.[17] This power calculation includes a significance level α = 0.05, a power β = 0.10 and a relative non-inferiority margin of δ = 0.90. Proving clinical superiority did not appear to be crucial to potential users of the tool in chapter 4 and so to avoid excessive data use the superiority margin was set to γ = 1.00. Although this non-inferiority margin is relative, due to the high NPV anticipated for CLC it will be similar and no larger than an absolute equivalent. This formed the rationale for the application of 10% non-inferiority margins applied in comparable studies using absolute rather than relative outcome measures.[12, 19] In evaluating this non-inferiority margin it is helpful to remember that the least desirable outcome of a false negative (FN) would yield a two or four week delay for the next planned treatment rather than treatment cessation and that 22% of patients are estimated to experience more than four weeks of delay to treatment in a year in current consultant-led care.[14] As such the null hypothesis (of inferiority) will be rejected if the lower confidence limit for the relative NPV of AI-enabled reports compared to CLC reports is greater than 0.90.[17] This will be visually presented with 2-sided 95% confidence interval (CI) (Figure 25).

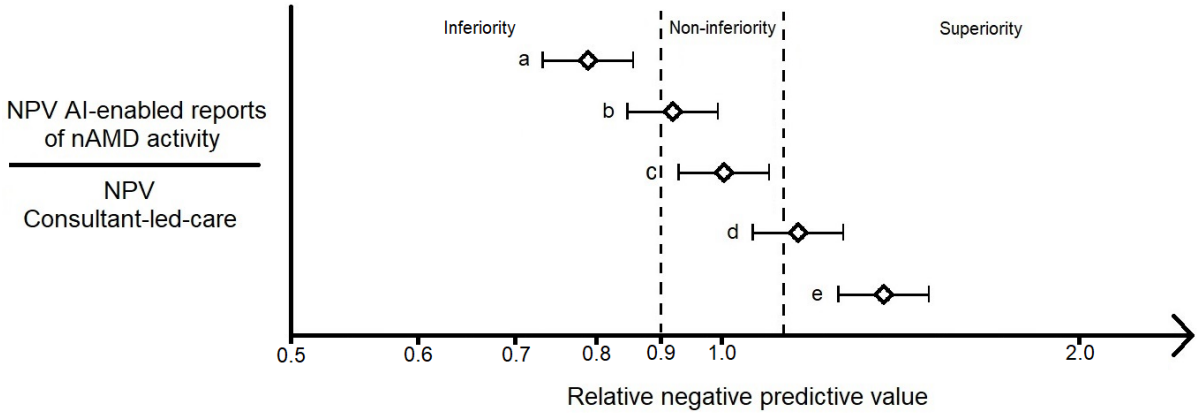


Figure 25.Forest plot template for the relative negative predictive value (NPV) of artificial intelligence (AI)-enabled reports of neovascular age-related macular degeneration (nAMD) disease activity versus judgements from consultant-led-care relative to an enhanced reference standard from Moorfields Reading Centre. The clinical non-inferiority and superiority margins are marked by dashed vertical lines on a logarithmic scale at 0.90 and 1.11 respectively. Potential outcomes for the non-inferiority test include scenario (a) AI-enabled reports are inferior to judgements from consultant-led-care; (b) Non-inferiority of AI-enabled reports to judgements from consultant-led-care is not demonstrated; (c) AI-enabled reports are non-inferior to judgements from consultant-led-care; (d) AI-enabled reports are non-inferior to judgements from consultant-led-care but not superior; (e) AI-enabled reports are superior to judgements from consultant-led-care.

Whilst the dataset will offer other important exploratory insights, the potential impact of the primary outcome on the translation of this AI-enabled decision tool and threat to the study's feasibility from ambitions outside this scope, has prevented any plans to proactively power the sample size for secondary outcomes.

5.5.2 Sampling method

The EMR at NuTH was searched to identify clinic visits where individuals received anti-VEGF IVT to treat nAMD. Patients at NuTH are treated under three different regimens dependent upon the length of their diagnosis and joint decisions between clinicians and patients: loading, treat-and-extend and pro-re-nata (Table 11). The planned treatment intervals vary between 4 and 16 weeks. During the period through which data were collected aflibercept and ranibizumab were the anti-VEGF treatments in use for nAMD.

Table 11 Summary of neovascular age-related macular degeneration (nAMD) treatment at Newcastle upon Tyne Hospitals NHS Foundation Trust. IVI=Intravitreal Injection, VEGF = Vascular Endothelial Growth Factor

Loading protocol: Starts at diagnosis of nAMD and consists of three anti-VEGF IVIs at four-week intervals (visits two and three do not include consultation or imaging), followed by IVIs at eight-week intervals, or less if signs of disease activity persist, for the remainder of the year.

Treat-and-extend (TEX) protocol: Starts one year after treatment initiation and dictates that the interval of anti-VEGF IVIs is increased in two-week increments until evidence of disease activity is noted at which point the treatment interval is reduced. If extension beyond a certain interval is noted to result in observable disease activity a (unspecified) number of times, then that interval ceases to be modified.

Pro re nata (PRN) protocol: Initiated as a joint decision for patients who appear to have little or no disease activity having been on one of the other two protocols. Here it is not assumed that an IVT will be given at each review, but only if evidence of disease activity is noted. The observation of returning disease activity may also lead to the return to one of the other two protocols.

From the dataset of potentially eligible visits (N=70,884), computer-generated randomised numbers were assigned to list patient EMR files to screen against eligibility criteria (Table 12). These criteria exclude clinical visits that took place outside current treatment and OCT imaging protocols and after treatment decisions may have been influenced by the Covid-19 pandemic. They also exclude visits conducted under the pro re nata (PRN) treatment protocol. This is because all visits on the PRN protocol meeting the inclusion criterion for a same day anti-VEGF IVT must have been judged by the clinician to show disease activity and would therefore not be representative of PRN clinic visits generally. To maximise their relevance to the research question alongside feasibility and rigour within a complex real-world dataset, the eligibility criteria (Table 13) and systematic screening approach through which they will be applied was iteratively designed and trialled by collaborators with clinical, operational and statistical expertise inside and outside of NuTH.

Table 12. AMD = Age-related macular degeneration, *IVT* = Intravitreal Treatment, VEGF = Vascular Endothelial Growth Factor, nAMD = neovascular Age-related macular degeneration, TEX = Treat and Extend, VA = Visual Acuity, OCT = Optical Coherence Tomography

Step	Question	Action if…	
		Yes	No
1	Does this eye have no retinal diagnosis beside AMD or is it enrolled in a study?	Go to step 2	Reject this patient
2	Is this visit more than 10 weeks after the eye's first IVT?	Go to step 3	Switch to the next visit, go to step 2

3	Does this visit involve anti-VEGF treatment for nAMD?	Go to step 4	Switch to the next visit, go to step 2
4	Is this visit conducted under the loading or TEX protocols?	Go to step 5	Switch to the next visit, go to step 2
5	Are the VAs of interest free from the influence of other interventions?	Go to step 6	Switch to the next visit, go to step 2
6	Does this visit have an accompanying consultation recorded?	Go to step 7	Switch to the next visit, go to step 2
7	Is the treatment interval stated?	Go to step 8	Switch to the next visit, go to step 2
8	Is there a VA available for this visit and the prior?	Go to step 9	Switch to the next visit, go to step 2
9	Are there co-located 25 slice fovea-centred OCTs available for this visit and the prior?	Collect data	Switch to the next visit, go to step 2

Having reached consensus on the eligibility criteria and the screening approach, a single researcher with 9 years of clinical experience at NuTH (JH) performed data collection, to support a consistent recruitment approach grounded in fluency with local clinical and digital practices.

Table 13. Eligibility criteria for patients (in bold) and clinic visits. nAMD = Neovascular age-related macular degeneration, IVT = Intravitreal Treatment, NuTH = Newcastle upon Tyne Hospitals, OCT = Optical coherence tomography

Inclusion criteria

- **Eye diagnosed with nAMD**
- **One or more prior anti-VEGF IVT at NuTH**
- Co-located 25 slice, fovea centred OCT imaging available for both the included and prior visits
- Clinic visit note states intended IVT interval
- Clinic visit included same-day IVI

Exclusion criteria

- **Retinal diagnosis other than nAMD in included eye**
- Visit before 2016
- Visit during or after March 2020
- Visits conducted under the pro-re-nata treatment protocol

5.5.3 Data collection and processing

For each included clinic visit the following data was recorded in the NuTH computer environment to characterise the dataset:

- Anonym for the individual
- Eye laterality
- Sex
- Self-reported ethnicity
- Home address postal code
- Individual's age at that visit

The following was recorded to send anonymised to Moorfields Reading Centre, to produce a report of disease stability or activity to act as an enhanced reference standard for each visit. This information also facilitated more meaningful post-hoc error analysis to explore the mechanisms of failure which the AI-enabled tool may exhibit. The findings from these analyses are a secondary outcome of the study and will help to delineate any groups of

cases for which the tool's performance needs to be monitored and improved in further work, or for which only clinician judgements should be applied. This list was developed through additions to a proforma from a recent exemplar protocol:[12]

- OCT and VA for that visit and the prior
- Type of VA (best corrected, pin hole or unaided)
- Presence or absence of evolving macular haemorrhage being recorded

The following data was recorded from each CLC visit to assess the primary and secondary outcomes and will not be sent to Moorfields Reading Centre:

- Judgement of disease activity or stability
- Planned interval to next IVI
- Professional group conducting the consultation
- Treatment protocol the visit was conducted under (Table 11)
- VA for the fellow eye at that visit
- Time since first nAMD treatment at that visit
- Total IVIs for nAMD in that eye up until that visit
- Observed interval since that eye's last IVI
- Time since increasing disease activity was last observed
- Treatment interval associated with that observation
- Mention of shared decision making in EMR entry

Separately, the present and prior pairs of OCT images relating to the same clinic visits will be transferred to Moorfields Eye Hospital NHS Foundation Trust for AI-enabled retinal segmentation.[13] The differences in retinal tissue volumes will be used in a rule-based decision tree to produce an AI-enabled binary report of disease activity or stability for each included visit.

5.5.4 AI-enabled decision tool

The intervention to be tested on retrospective OCT imaging data is a deep learning tool with a U-net architecture, called OCTane, with previously published details of training and validation.[13] OCTane can produce volume quantification for the neurosensory retina, retinal pigment epithelium, fibrovascular pigment epithelium detachment, drusenoid

pigment epithelium detachment, SRHM, SRF, IRF, posterior hyaloid, epiretinal membrane and serous pigment epithelium detachment. At the time of writing OCTane is not regulated as SaMD and so this functionality is only available in a research setting. Notably, these scalar outputs of various tissue volumes are not directly actionable as judgements of disease activity or management recommendation. To add this functionality, an initial decision tree will report disease activity when the volumes of SRF and/or IRF increase between the current and prior visit OCTs. This decision logic is based on a recent consensus from UK medical retina experts on treat-and-extend protocols for nAMD.[9] The exact tissue group contributors and decision thresholds for inter-visit changes in each of these tissue groups will be iterated upon using an embedded pilot dataset described further below. This binary output was preferred over a scalar recommendation of treatment interval to preserve the tools' value across different treatment protocols and therefore clinical contexts (Table 11).

5.5.5 Outcomes measures

Over-treatment marginally increases the cumulative risk of IVT complications and the cost to the provider, but justifiably the main concern of patients and carer participants in chapter 4 and elsewhere in the literature was sight loss through under-treatment.[14, 15] Consequently, the probability of AI-enabled reports of disease stability being correct relative to judgements made in real-world consultant-led care has been taken as the most clinically relevant measure of diagnostic accuracy. This has led to a non-inferiority design with the relative negative predictive value (NPV) as the primary outcome.

Secondary outcomes will be:

1. A comparison of other standard diagnostic accuracy metrics between AI-enabled reports of disease activity (across various thresholds) and judgements from CLC accompanied by confusion matrices (Table 10)
2. A comparison of the diagnostic accuracy of the five healthcare professional groups conducting consultations in CLC (nurses, optometrists, ophthalmology specialty trainees, medical retina sub-specialty fellows, medical retina sub-specialty consultants)
3. A comparison of the treatment intervals recommended in real-world consultant-led care and the treatment intervals that would be derived from AI-enabled reports of disease activity given the treatment protocol
4. A case-by-case exploration of false-positive and false-negative reports of disease activity from the AI-enabled decision tool and consultant-led care

5.5.6 Data analysis

The relative NPV of AI-enabled reports / judgements from CLC will be calculated with 95% CIs to see if the inferiority, non-inferiority or superiority of AI-enabled reports of disease activity can be established (Figure 25). A similar calculation will also be applied to relative positive predictive value (PPV) to aid in the interpretation of the primary outcomes. Both of these calculations depend on the derivation of σ^2_N from the proportions of different categories of agreement and disagreement between the AI-enabled reports, CLC judgements and Moorfields Reading Centre judgements and the estimates of NPV for the AI-enabled reports and CLC judgements.[17]

$$\sigma^2_N = 1 / [(p_2 + p_4)(p_3 + p_4)] \times [NPVX_2(-p_3 + p_4 - 2(p_4 + p_8) NPVX_1) + (p_2 + p_3) - NPVX_1(p_2 - p_4)]$$

$$\sigma^2_P = 1 / [(p_5 + p_7)(p_5 + p_6)] \times [p_6(1 - PPVX_2) + p_5(PPVX_2 - PPVX_1) + 2(p_7 + p_3)PPVX_1 \times PPVX_2 + p_7(1 - 3PPVX_1)]$$

This can then be applied with the value of α (0.05) and the size of the final dataset (N) to derive the CIs for both the estimates of relative NPV (rNPV) and relative PPV (rPPV).[17]

$$rNPV \pm 95\%\ CI = e^{[ln\ rNPV \pm z_{1-\alpha} \sqrt{(\sigma^2_N/N)}]}$$

$$rPPV \pm 95\%\ CI = e^{[ln\ rPPV \pm z_{1-\alpha} \sqrt{(\sigma^2_P/N)}]}$$

For secondary outcomes, diagnostic accuracy statistics (presented as a proportion, p, estimated from a sample of size N) for each group will be reported descriptively with 95% CIs for each group, along with confusion matrices. These CIs will be derived by the Clopper-Pearson method.[20] This was selected over the simple normal approximation (Wald interval) as many of the estimates approach 100% which would lead to implausible upper CIs being calculated in excess of 100% with the Wald interval. The Clopper-Pearson method also provides reliably conservative CIs, minimising the risk of type 1 error in interpreting the findings which is thought to be a greater risk with the Wald interval.[21]

$$Lower\ CI = [1 + 3.247 \times ((1-p) + 1/N)/p]^{-1}$$

$$Upper\ CI = [1 + ((1-p)/((1/N + p) \times 2.2882)]^{-1}$$

Clinical and imaging data from cases of false positives (FPs) and FNs of the AI-enabled reports were reviewed by clinical members of the team, supported by the AI development team where necessary, to try to understand the mechanisms of AI-enabled decision tool failures.

5.6 Results

5.6.1 Pilot dataset and power calculation

Applying the systematic sampling strategy (Table 12) to sequentially randomly select cases from the NuTH dataset of 70,884 clinic visits produced by an EMR search presented a lower prevalence of disease stability (negatives) in clinic appointments than the initial estimate of 79%. Consequently, random sampling continued beyond the initially forecast 127 cases to acquire 135 eligible cases, 102 of which (75.6%) were judged to demonstrate disease stability by CLC. These cases were then processed by Moorfields Reading Centre where 104 of the 135 cases (77.0%) were judged to demonstrate disease stability (Table 14).

Table 14. Judgements of disease activity from consultant-led-care (CLC) and Moorfields Reading Centre (MRC) in the pilot dataset of 135 clinic visits. TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

	MRC +ve	MRC -ve	Totals
CLC +ve	TP 14	FP 19	33
CLC -ve	FN 17	TN 85	102
Totals	31	104	135

These data provided the estimate of NPV for CLC (X_2) required for the power calculation and the PPV for balanced interpretation; NPV = 83.3% (95% CI 70.8-95.9), PPV = 42.4% (95% CI 34.1-50.7). OCT scans for the same cases were then analysed by OCTane to produce scalar outputs for tissue volumes in each OCT scan in the pilot dataset. These absolute tissue volumes were then converted into changes in tissue volume between the study visit and the prior visit by simple subtraction. Three different rule sets, reflecting a range of clinical rationales applied in current practice, were then applied to these scalar outputs to produce binary judgements of disease activity and NPV and PPV estimates were then derived for each. All 3 rule sets provided estimates of NPV that were greater than that of CLC and so the rule set which did not appear to confer a reduction in PPV relative to CLC was selected for the power calculation. This aimed to demonstrate non-inferiority for an AI-enabled decision tool with regard to under-treatment without having to tolerate more over-treatment in a potential service. Taking D as the Moorfields Reading Centre Judgment, X_1 as the AI-enabled judgement, X_2 as the CLC care judgement, 1 to represent a judgement of disease activity and 0 to represent a judgment of disease activity this returned the following 8 disagreement types quantified in Table 15.

Table 15. Types of disagreement between judgements of disease activity in the pilot dataset made by Moorfields Reading Centre (D), rule set 1 overlaid on OCTane outputs (X_1) and consultant-led-care (X_2). Disagreements are expressed across 8 categories as integers (n) and proportions (p).

	D = 0 (104 cases)		D = 1 (31 cases)	
	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	$n_1 = 5$ $p_1 = 0.04$	$n_2 = 25$ $p_2 = 0.19$	$n_5 = 11$ $p_5 = 0.08$	$n_6 = 12$ $p_6 = 0.09$
$X_1 = 0$	$n_3 = 14$ $p_3 = 0.10$	$n_4 = 60$ $p_4 = 0.44$	$n_7 = 3$ $p_7 = 0.02$	$n_8 = 5$ $p_8 = 0.04$

Proportions for each of these different types of disagreement could be carried forward with the estimate of 83.3% for the NPV of CLC and the predetermined values for α, β, γ and δ to apply the sample size calculation for non-inferiority. Rounding to the nearest integer this returned a recommended sample size of 262 cases to test the non-inferiority of the rNPV of AI-enabled judgements of disease activity compared to CLC. Using the same sampling approach, a further 127 cases were then extracted and processed by Moorfields Reading Centre and OCTane to combine with the initial 135 cases.

5.6.2 Final dataset

Prior to anonymising and exporting the final dataset, a final round of data cleaning was performed against the EMR. This led to the alteration of some of the data held in initial export of the pilot dataset. This included the primary outcome measure in a handful of cases. This was largely due to comparing treatment interval recommendations from CLC with the actual observed interval since the last treatment, rather than the interval that was recommended at the prior visit. Discrepancies here were because of the frequency of appointment delays in real-world care. This is quantified and discussed further below.

5.6.2.1 Dataset characteristics

In applying the systematic sampling strategy to identify 262 eligible visits, 159 were assessed as ineligible and for 84 of eyes which the initially identified visits focused on, it was not possible to identify an eligible visit. The frequency of the different reasons for these ineligibilities is detailed in Table 16.

Table 16. Type and frequency of visit exclusions during screening; IVT = intravitreal treatment, nAMD = neovascular age-related macular degeneration, VA = visual acuity, OCT = optical coherence tomography

Step number	Exclusion criteria	Ineligibilities for initially identified clinic visit n=159 (%)	No eligible case available for eye n=82 (%)
1	The visit is less than 10 weeks after the eye's first IVI	1 (0.6)	0 (0.0)
2	The eye has a retinal diagnosis beside nAMD or is enrolled in a study	44 (27.7)	42 (51.2)
3	The visit does not involve anti-VEGF treatment for nAMD	1 (0.6)	1 (1.2)
4	The VA measurements of interest are likely influenced by other interventions	3 (1.9)	1 (1.2)
5	There is no consultation at this visit	52 (32.7)	15 (18.3)
6	The clinician does not state the treatment interval they intend	1 (0.6)	0 (0.0)

7	VA measurements for the visit and the prior one are unavailable	0 (0.0)	0 (0.0)
8	Co-located 25 slice fovea-centred OCTs are not available for the visit and the prior one	29 (18.2)	9 (11.0)
9	The visit is under the PRN protocol	28 (17.6)	14 (17.1)

The 262 eligible visits related to 238 distinct individuals, with 24 individuals contributing two eligible visits to the dataset. Of the data extracted for the eligible visits age, sex, ethnicity, drug used for IVI, VA at diagnosis and income deprivation affecting older people index (IDAOPI) decile were available for the full dataset. Scalar variable (Table 17) means were compared for visits that were sampled (n=262) and not (n=30557) with two-sided independent t-tests and were found to be statistically different but clinically equivocal for age (79.8 (95% CI 78.9 – 80.7) vs 78.5 (78.4 – 78.5), p=0.005) and statistically equivocal for baseline VA (60.1 (58.4 – 61.7) vs 59.7 (59.5 – 59.8)) and IDAOPI decile (5.2 (4.8 – 5.5) vs 5.3 (5.3 - 5.4)). Categorical variable (Table 18) proportions were compared using Chi-squared tests and were found to be equivocal for sex (37.7% male (37.1 – 38.2) vs 38.7% (32.8 – 44.6), p=0.74), ethnicity (90.1% recorded as white (89.8 – 90.4) vs 92.7% (89.6 – 95.9), p=0.69) and drug (95.3% aflibercept (95.0 – 95.5) vs 98.5% (97.0 – 100.0), p=0.114).

Table 17. Categorical data characterising the final dataset of 262 eligible clinic visits. SAS = Specialty doctors and Associate Specialists, TEX = Treat and Extend

Characteristic	Categories	n (N=262)	%
Concurrent bilateral treatment	Yes	77	70.6
	No	185	29.4
Sex	Male	101	38.5
	Female	161	61.5
Ethnicity	British	243	92.7
	Pakistani	1	0.4
	Not stated	18	6.9
Study eye laterality	Left	122	46.6
	Right	140	53.4
Treatment drug	Aflibercept	238	90.8
	Ranibizumab	24	9.2
Diabetic status	Not diabetic	154	58.5
	Type 2	45	17.2

	Diabetic – type unknown	4	1.5
	Status unknown	59	22.5
Consulting clinician	Nurse practitioner	78	29.8
	Optometrist/orthoptist	19	7.3
	Ophthalmology trainee	15	5.7
	Ophthalmology fellow	77	29.4
	Ophthalmology consultant/SAS	73	27.9
Protocol	Loading	53	20.2
	TEX	209	79.8

These characteristics (most notably ethnicity and diabetic status) depended on the quality of data recording in the EMR, but beside entries of 'unknown' there was no missing data. Although it was not possible to clarify the ethnicity of the 18 patients with 'not stated' recorded, other personal data such as name and religion did not suggest that the ethnic make-up of this sub-cohort was different to the other 244 cases.

Table 18.Scalar or ordinal data characterising the final dataset of 262 eligible clinic visits. VA= visual acuity, nAMD = neovascular age-related macular degeneration, IVT = intravitreal treatment, IQR = Interquartile Range

Characteristic	Median	IQR
Deprivation Affecting Older People Index national decile	5	3-7
Age at visit	81	76-85
Study eye VA at nAMD diagnosis	62	51-70
Number of prior IVT to study eye	11	7-17
Observed interval since prior visit	8	7-10
Interval planned at prior visit	8	6-8
Study eye VA at visit	67	55-74
Contralateral eye VA at visit	69	44-77

5.6.2.2 Primary outcome

In the 262 clinic visits, Moorfields Reading Centre identified 71 cases of disease activity, estimating a prevalence in the clinic population of 27.1%. Relative to this reference standard, CLC exhibited a NPV of 80.8% (95% CI 71.3-90.4) and a PPV of 42.2% (95% CI 30.03-54.2) (Table 19). This included 33 sight-threatening 'false negatives' where CLC

recommended maintenance or reduction of the treatment interval whilst the reference standard suggested that shortening of the treatment interval would have been appropriate.

Table 19. 2x2 table for consultant led care (CLC) judgements of disease activity compared to the reference standard provided by Moorfields Reading Centre (MRC) for the final dataset. TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

	MRC +ve	MRC -ve	Totals
CLC +ve	TP 38	FP 52	90
CLC -ve	FN 33	TN 139	172
Totals	71	191	262

Applying rule set 1 to OCTane segmentation outputs, which supported the power calculation and only regarded any increase of IRF as evidence of disease activity, produced an NPV estimate of 85.5% (77.0-94.0) for AI-enabled judgements of disease activity and a PPV estimate of 44.5% (95% CI 32.5-56.6) (Table 20). This included 22 sight-threatening 'false negatives' where rule set 1 recommended maintenance or reduction of the treatment interval whilst the reference standard suggested that shortening of the treatment interval would have been appropriate. This estimate gave a rNPV of 1.06 (0.855/0.808) and a rPPV of 1.05 (0.445/0.422).

Table 20. 2x2 table for AI-enabled judgements of disease activity using rule set 1 (R1) compared to the reference standard provided by Moorfields Reading Centre (MRC) for the final dataset. TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

	MRC +ve	MRC -ve	Totals
R1 +ve	TP 49	FP 61	110
R1 -ve	FN 22	TN 130	152
Totals	71	191	262

To calculate the 95% CIs and complete the non-inferiority test the formula listed in 4.5.6 was used.[17] Taking D as the Moorfields Reading Centre judgment, X_1 as the AI-enabled

judgement under rule set 1, X_2 as the CLC care judgement, 1 to represent a judgement of disease activity and 0 to represent a judgment of disease activity this returned the disagreement types quantified in Table 21.

Table 21. Types of disagreement between judgements of disease activity in the final dataset made by Moorfields Reading Centre (D), rule set 1 applied to OCTane outputs (X_1) and consultant-led-care (X_2). Disagreements are expressed across 8 categories as integers (n) and proportions (p).

	D = 0 (191 cases)		D = 1 (71 cases)	
	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	$n_1 = 14$ $p_1 = 0.05$	$n_2 = 47$ $p_2 = 0.18$	$n_5 = 30$ $p_5 = 0.11$	$n_6 = 19$ $p_6 = 0.07$
$X_1 = 0$	$n_3 = 38$ $p_3 = 0.15$	$n_4 = 92$ $p_4 = 0.35$	$n_7 = 8$ $p_7 = 0.03$	$n_8 = 14$ $p_8 = 0.05$

A 95% CI for rNPV of 0.99 – 1.13 was derived, the lower bound of which is greater than the pre-determined clinical margin for non-inferiority (δ) of 0.9. It is therefore possible to reject the null hypothesis and accept that these AI-enabled judgements of disease activity are non-inferior to judgements from CLC with regard to NPV (i.e. sight-threatening undertreatment). The 95% CI for rPPV was broader at 0.84 – 1.32 but suggests that a lower PPV should not be expected from these AI-enabled judgements (i.e. costly overtreatment) should they be implemented.

5.6.2.3 Secondary outcomes

5.6.2.3.1 Diagnostic accuracy statistics across various rule sets

Given the pragmatic goal of this study and the closeness with which the dataset represents real-world practice, NPV and PPV were selected as clinically intuitive outcome measures. However, sensitivity, specificity and likelihood ratios are common measures of diagnostic accuracy which add other nuances for interpretation. Similarly, to respect the a-priori protocol of the study, few rule sets were explored and the one which appeared to perform most favourably on the pilot dataset was carried through to the power calculation and primary outcome test. Given the novelty of designing treatment rationales on objective continuous outputs of tissue volumes, rather than the subjective impression of clinicians, it may well be that other decision thresholds offer greater performance. To address both of these opportunities a wider range of diagnostic accuracy statistics were calculated (Table 23) for a total of seven different rule sets (Table 22).

Table 22. Rule sets overlaid on OCTane outputs to derive AI-enabled judgements of disease activity. IRF = intraretinal fluid, SRF =subretinal fluid, SHRM = subretinal hyper-reflective material

Rule set	Disease activity if…
1	any increase in IRF
2	any increase in IRF or SRF
3	any increase in IRF, SRF or SRHM

4	>1% increase in neurosensory retina
5	>10% increase in IRF
6	>10% increase in IRF or SRF
7	>10% increase in IRF, SRF or SRHM

Each of these rule sets remains rooted in the rationale applied in clinical decision making.[9]

Table 23. Diagnostic accuracy statistics with 95% confidence intervals for consultant-led-care (CLC) and 7 different rulesets (Table 22). LR+ = Positive likelihood ratio, LR- = Negative likelihood ratio, NPV = negative predictive value, PPV = positive predictive value

Rule set	NPV	PPV	Sensitivity	Specificity	LR+	LR-
CLC	80.8% 55.7-90.7	42.2% 18.1-63.2	53.5% 25.6-73.0	72.8% 44.7-86.0	1.97	0.64
1	85.5% 63.5-93.2	44.5% 19.6-65.2	69.0% 39.6-83.9	68.1% 39.2-83.1	2.16	0.46
2	94.9% 82.7-97.7	40.2% 17.0-61.0	93.0% 77.2-96.8	48.7% 22.4-68.7	1.81	0.14
3	95.9% 84.5-98.2	36.2% 14.8-56.8	95.8% 84.0-98.1	37.2% 15.3-57.9	1.52	0.11
4	80.7% 55.6-90.6	56.4% 27.6-75.3	43.7% 18.9-64.7	87.4% 67.3-94.1	3.47	0.64
5	86.0% 64.4-93.4	46.7% 20.9-67.1	69.0% 39.6-83.9	70.7% 42.2-84.7	2.35	0.44
6	95.3% 83.8-97.9	42.3% 18.3-63.0	93.0% 77.2-96.8	52.9% 25.5-72.2	1.97	0.13
7	96.3% 85.7-98.4	37.6% 15.5-58.3	95.8% 84.0-98.1	40.8% 17.4-61.5	1.62	0.10

Comparing the primary outcome of NPV across the rule sets, the only option which appears to offer further gains in NPV whilst not compromising on the PPV of CLC is rule set 6 (Figure 26). If binary decisions from rule set 6 are analysed across the dataset, the rNPV against CLC is 1.18 (1.09-1.27) whilst the rPPV is 1.00 (0.82-1.23). This appears to represent AI-enabled decisions which have a NPV that is statistically (though perhaps not clinically) superior to CLC, without suggestion that PPV would be lowered.

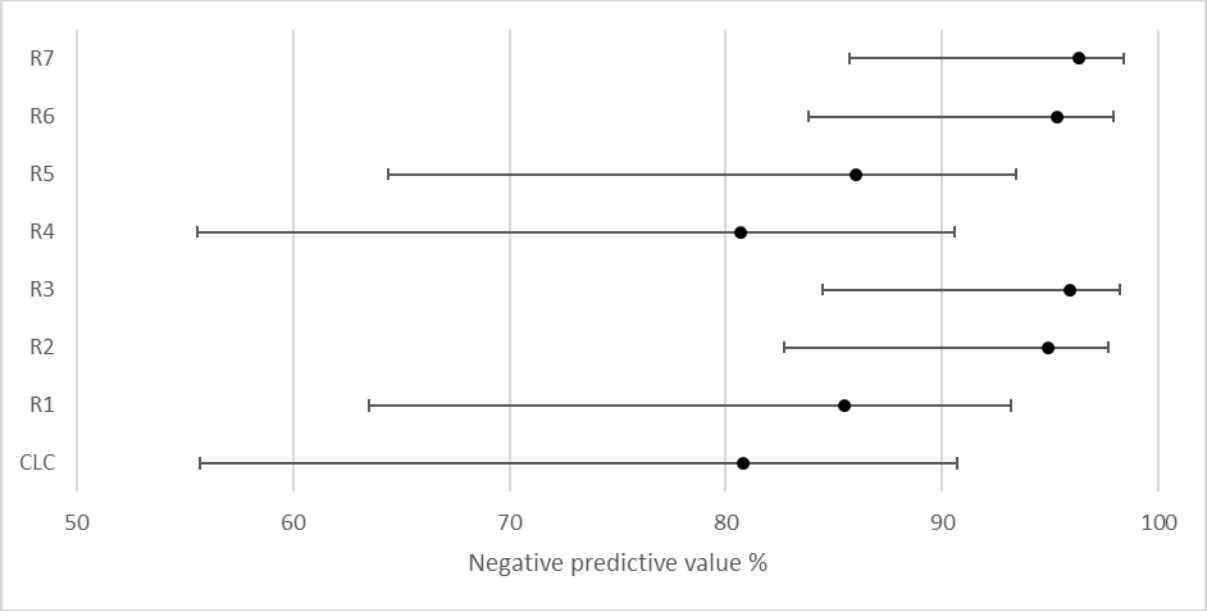


Figure 26. Forest plot comparing the negative predictive value of judgements of disease activity made by applying rule sets 1 - 7 (R1 – R7) to OCTane outputs with consultant-led care (CLC). Error bars display 95% confidence intervals calculated using the Clopper-Pearson method.[20]

5.6.2.3.2 Diagnostic accuracy across professional groups

The professional group which conducted the consultations in CLC was recorded for each visit (Table 17). These groups still hold a reasonable degree of heterogeneity within them but mirror the ways in which staff are categorised in clinical practice. The most heterogenous group was that of fellows, who conducted 77 (29.4%) of the consultations. This group includes doctors yet to specialise with as little as 2 years of experience in clinical practice, UK trained ophthalmologists with consultant level experience and ophthalmologists trained abroad with widely varied experience and future career intentions. Due to small sample sizes some of these CIs are very wide (Table 17).

Table 24. Negative predictive value (NPV) and positive predictive value (PPV) of consultant-led-care judgements of disease activity by professional group. SAS = Specialty and Associate Specialist doctors, CI = confidence interval

	n	NPV	95% CI	PPV	95% CI
Rule set 6	262	95.3%	83.8 – 97.9	42.3%	18.3 – 63.0
Consultants/SAS	73	88.2%	66.4 – 94.6	50.0%	22.0 – 71.4
Nurse practitioners	78	85.2%	61.2 – 93.1	58.3%	28.2 – 77.4
Ophthalmology trainees	15	83.3%	50.7 – 92.6	100.0%	48.0 – 100.0
Retina fellows	77	71.1%	41.3 – 85.3	18.8%	6.4 – 38.1
Optometrists/ orthoptists	19	60.0%	27.0 – 80.0	44.4%	17.0 – 69.6

As a further subgroup analysis, the rNPV and rPPV was recalculated for rule set 6 compared to F2F decisions made by consultants or specialty doctor and associate specialist (SAS) doctors (n=73). This returned a rNPV of 1.07 (95% CI 1.01 – 1.13) and a rPPV of 0.77 (95% CI 0.63 – 0.93).

5.6.2.3.3 Comparison of treatment burden from CLC and AI-enabled decisions

Both the cost of service provision and patient satisfaction are related to the frequency of IVIs required by an individual. Diagnostic accuracy data and treatment protocols were used to model the consequences of different approaches to clinical decision making on IVT frequency. Judgements of disease activity were translated into recommended treatment intervals, knowing what the observed prior treatment interval had been and whether the patient was on the TEX or loading protocol (Figure 27Figure 27. Approach to convert judgements of disease activity into recommended treatment intervals.). To make these data more meaningful representatives of treatment burden they were converted to an annual IVT rate by dividing the recommended treatment interval in weeks by 52.

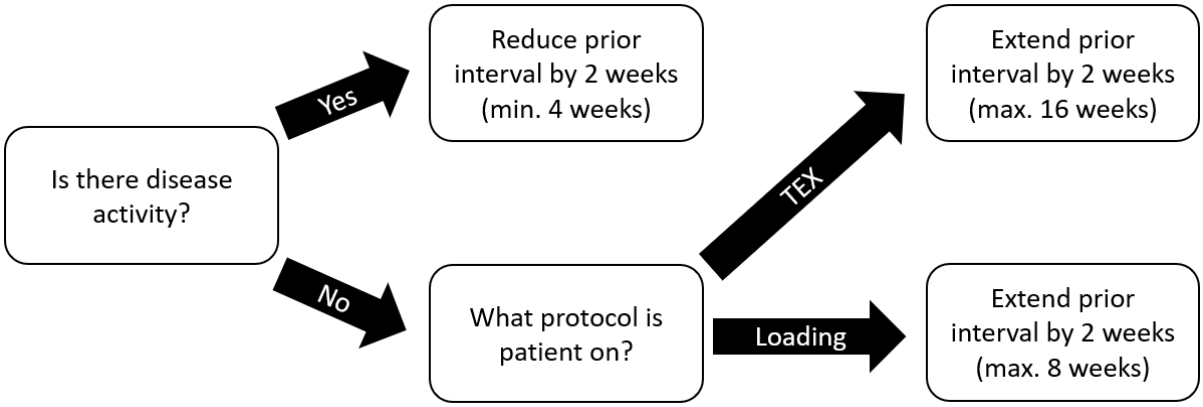


Figure 27. Approach to convert judgements of disease activity into recommended treatment intervals.

To validate this approach of simulating real-world care, the observed treatment interval recommendations from CLC were compared to those derived from the binary judgements of disease activity applied to Figure 27. The mean annual IVT rate observed from CLC was 6.98 IVIs/year (95% CI 6.73 – 7.23) which appeared equivocal to the mean rate of 7.08 IVIs/year (95% CI 6.76 – 7.39) derived using Figure 27. This equivalence on aggregate may be driven by patients on the TEX protocol however as the IVT rates from CLC judgements of disease activity for patients on the loading protocol appear to be greater, when derived from Figure 27, than they were observed to be from real-world recommendations.

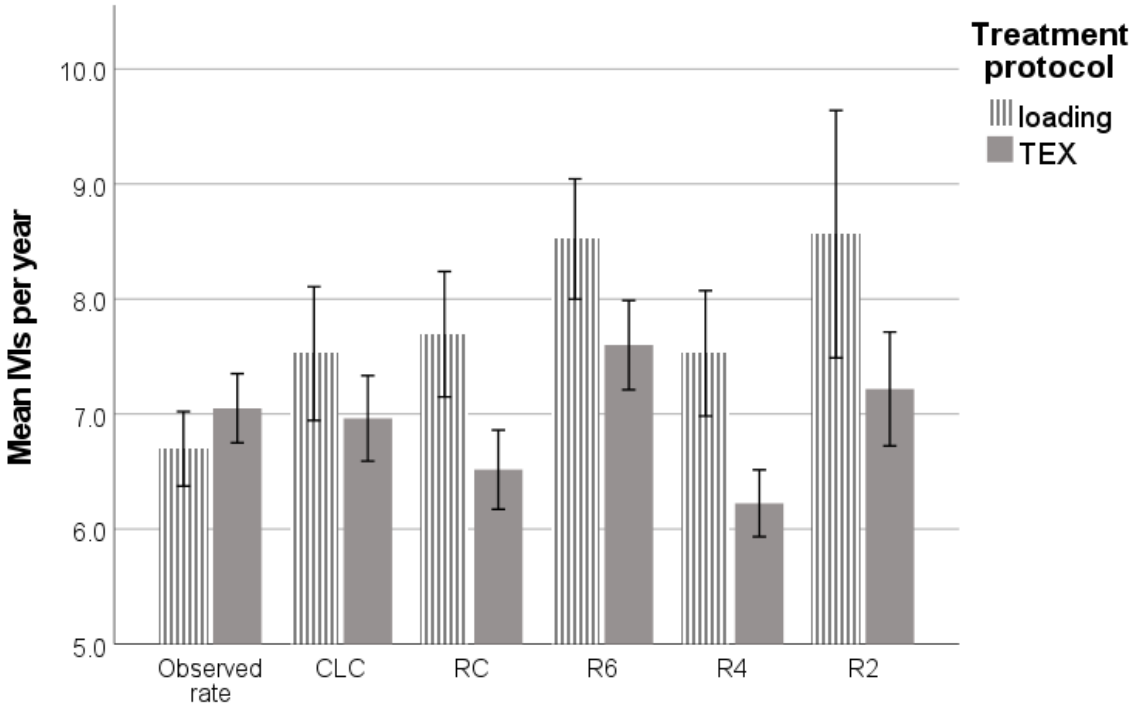


Figure 28. Intravitreal Injection (IVI) treatment rates from observed treatment interval recommendations and from binary judgements of disease activity from consultant led care (CLC), Moorfields Reading Centre (RC), and rule sets (R) 2,4 and 6.

Relative to the observed rate of IVIs in real-world clinics, those derived from Moorfield Reading Centre judgements of disease activity appeared equivocal. The IVT rates derived from rule sets 2 and 6 appeared 0.1 IVIs/year (95% CI 0.0-0.3) and 0.4 IVIs/year greater (95% CI 0.2-0.5) respectively, whilst the IVT rates from rule set 4 appeared 0.2 IVIs/year lower (95% CI 0.1-0.3).

5.6.2.3.4 Error analysis

In accordance with the study design, the enhanced reference standard from Moorfields Reading Centre is taken as the ground truth. As with any subjective clinical decision this is open to challenge. To evaluate the reference standard, CLC judgements of nAMD disease activity at the visit which followed the study visit and changes in VA at that subsequent visit were also recorded. For the 260 visits where a subsequent visit was available in the EMR, there was no significant difference (judged by 95% CIs) in the rates of subsequent disease stability (78.8% (64.6, 93.0) vs 79.7% (74.4, 85.1)) or VA change (-1.6 (-4.4,1.2) letters Vs -0.6 (-1.5, 0.43) letters) between study visits with false negative (FN) CLC judgements (n=33) and the other study visits (n=227) respectively. Across all 262 cases, CLC judgements were 67.6% accurate with 33 FNs and 52 FPs and judgements derived from rule set 6 were 64.1% accurate with 5 FNs and 89 FPs.

Independent t-tests and chi-squared tests were used to screen for any scalar or categorical variables which may be associated with FP or FN errors from CLC or R6 (see appendix). This screening exercise involved 60 statistical tests, so the p values should be interpreted with caution, however very few suggested potential unequal performance. For CLC, there was a weak suggestion that FPs may be more common in patients being treated at a lower frequency or on TEX rather than loading protocols (these characteristics overlap considerably clinically). It also aligns with observations from practice and EMR interrogation (JH) where clinicians' often exercise caution on the TEX protocol rather than aggressively

seeking treatment extension. For R6 there was no suggestion of unequal distribution of FNs between different groups, but as only 5 of 262 visits were R6 FNs this cannot provide strong assurance of equal performance. For FPs there was a very weak suggestion that R6 may be more likely to produce FPs for cases which have more recently demonstrated a recurrence of disease activity. This p value of 0.05 may well be a chance occurrence, but it may reflect the clinical rationale that OCT scans and the differences between them can be more challenging to interpret in the context of recent disease activity.

Imaging and segmentations were reviewed for either five randomly sampled cases or comprehensive sampling of cases from each of the six categories of disagreement between judgements of disease activity from Moorfields Reading Centre, CLC and rule set 6 (Table 25). Qualitative descriptions of the original OCT images and the OCTane segmentations in these cases are summarised in the sections below.

Table 25. The distribution of 262 cases across the eight different categories of disagreement. Red highlights errors by both rule-set 6 (R6) and consultant-led-care (CLC), yellow highlights an error by one of CLC or R6 and green highlights correct judgements from CLC and R6. MRC = Moorfields reading centre

	MRC -ve (191 cases)		MRC +ve (71 cases)	
	CLC +ve	CLC -ve	CLC +ve	CLC -ve
R6 +ve	24	66	36	30
R6 -ve	28	73	2	3

5.6.2.3.4.1 FPs for both CLC and R6 (n=24)

Of the 5 cases interrogated, 1 was clinically ambiguous. For CLC, the cause of FPs was not always clear, but sometimes appeared to be related to small drops in patient VA or apparent failure to appreciate the difference between the actual and intended prior treatment interval, e.g. recommending an 8 week interval because that was recommended last time although there was still disease stability after a delayed 10-week appointment. For R6 it seemed that it was susceptible to FPs when very small increases in SRF occurred on a baseline of little or no SRF on the prior scan. This readily crossed the 10% threshold of R6 even though such increases hold little clinical significance and were sometimes simply the result of minor segmentation errors. Suboptimal image quality often led to segmentation errors which could easily trigger a FP if it involved overestimation of IRF on the study scan or underestimation on the prior scan (Figure 29). The false attribution of outer retinal tubulation as SRF by both clinicians and OCTane also appears to have played a role in one case.

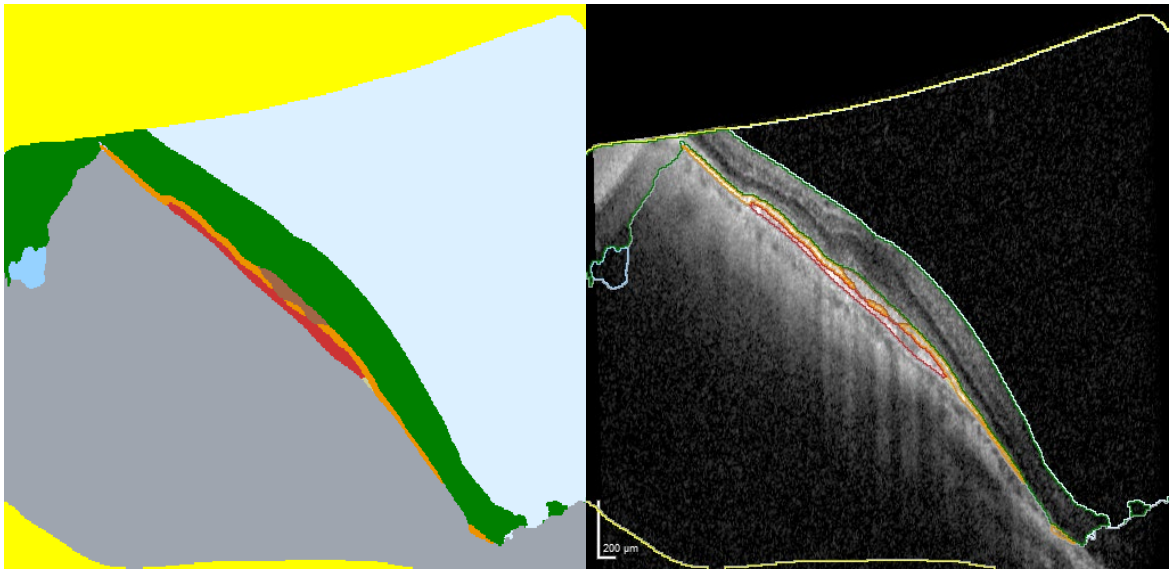


Figure 29. False positive from ruleset 6 where a mirror artefact (left-side) and low illumination (right-side) are associated with segmentation errors, including the false identification of intraretinal fluid (light blue).

5.6.2.3.4.2 True negatives for CLC and FPs for R6 (n=66)

All 5 of the cases sampled here appeared to share the same mechanism for the R6 FP, whilst CLC correctly identified disease stability. This was because of the proportional R6 threshold for disease activity regarding SRF. If a prior scan had little or no SRF, then even tiny volumes of isolated SRF (<1,000,000um^3) on the subsequent study scan would be enough to trigger a FP (Figure 30). This was particularly problematic because minor segmentation errors from OCTane where very small volumes are incorrectly assigned as SRF do not appear to be uncommon. Because OCTane was designed to process 50 B-scan volumes and NuTH captures 25 B-scan volumes routinely, a duplication step was part of pre-processing meaning that each B-scan was segmented twice. Interestingly, the erroneous attribution of SRF was often not repeated on the duplicate scan which may be indicative of low confidence from the segmentation model.

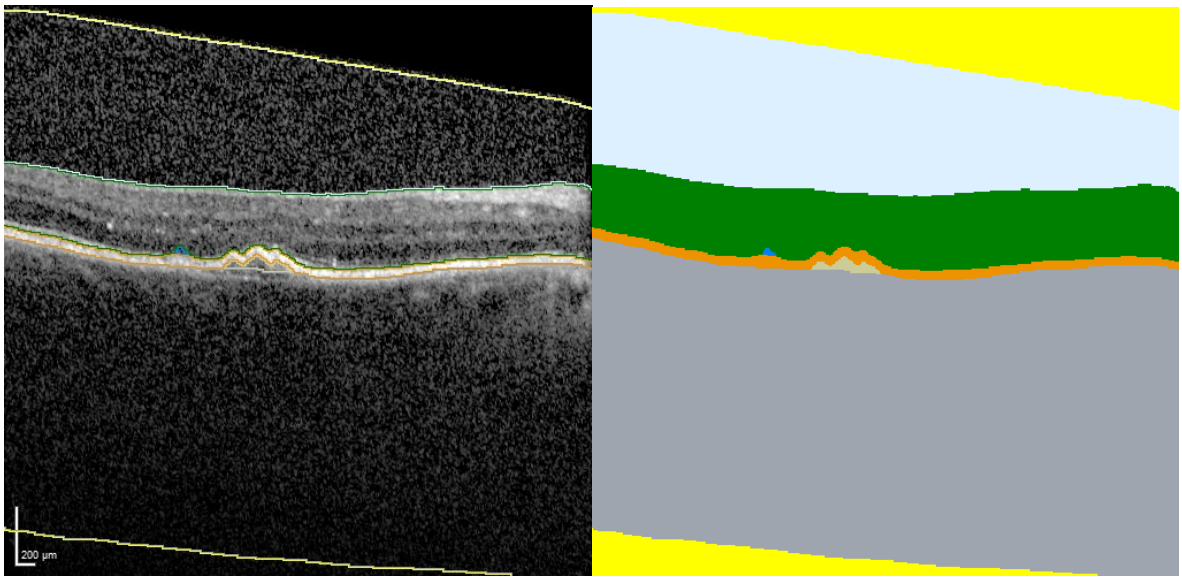


Figure 30. Small volumes of subretinal fluid (royal blue) segmented accurately or inaccurately by OCTane can lead to false positives.

5.6.2.3.4.3 FPs for CLC and true negatives for R6 (n=28)

FPs from CLC appeared to be most related to circumstances where clinicians perceived disease stability and sought to maintain the prior treatment plan (e.g. 8 weeks). The issue arose because they failed to account for a delay in the current visit which meant that the disease stability they were observing had followed a longer interval (e.g. 10 weeks). This meant that by maintaining the prior treatment interval (e.g. 8 weeks) they were in fact behaving as though they had observed disease activity. There were also some instances where there were subtle shifts in fluid volumes between scans which could easily be interpreted as greater or lesser without an objective tool.

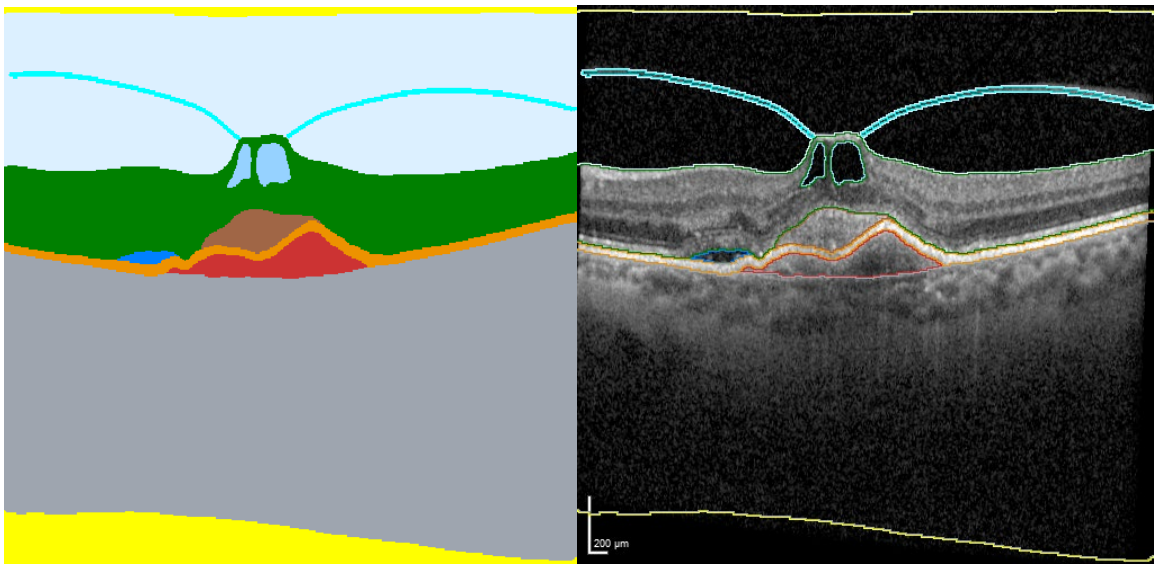


Figure 31. Accurate segmentation from OCTane despite a complex image involving traction between the posterior hyaloid (cyan) and the neurosensory retina (green), fibrovascular pigment epithelial detachment (red), subretinal hyper-reflective material (brown) and subretinal fluid (royal blue). Despite this accurate segmentation, if the intraretinal fluid associated with the VMT has increased between visits, neovascular age-related macular degeneration activity would have been inappropriately identified by rule set 6.

5.6.2.3.4.4 FNs for CLC and true positives for R6 (n=30)

Similarly to the CLC FPs above, some of the CLC FNs arose from cases where there were subtle shifts in fluid volumes. On one case reviewed, clear new SRF was not commented on or acted on in the treatment plan. This may have been an oversight, or may be reflective of some of the variation in approach between clinicians in the significance they place upon isolated SRF change. There were also examples where R6 had arrived at the correct outcome despite bad segmentation errors involving the false segmentation of large volumes of IRF at the study visit. This was a case with a large element of fibrovascular pigment epithelium detachment which also had genuine new IRF at the study visit (Figure 32). Had the segmentation error happened to have occur at the prior visit R6 would also have produced a sight-threatening FN.

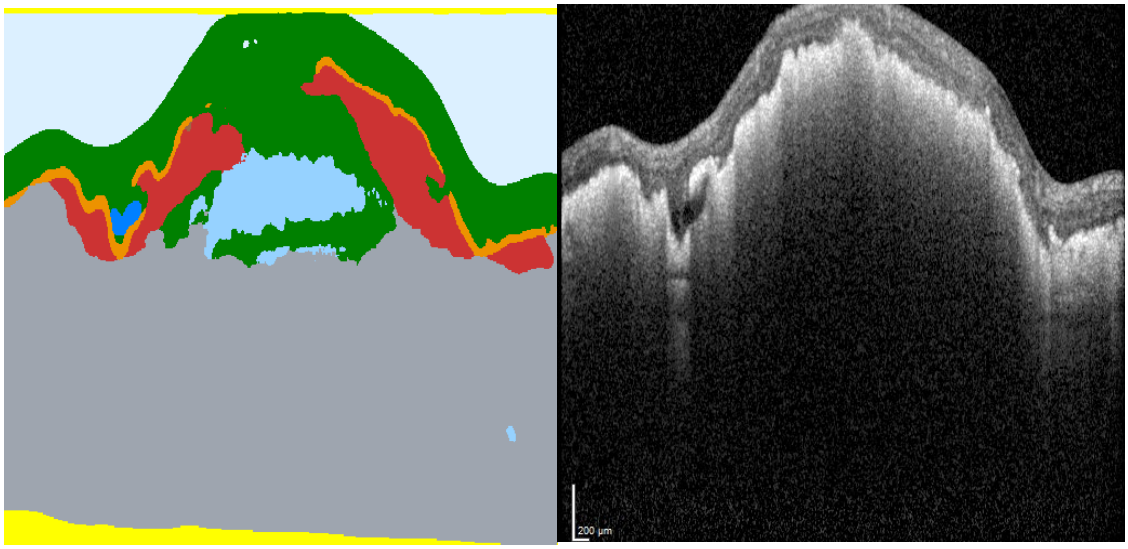


Figure 32. OCTane segmentation error, attributing a large cross-sectional area of fibrovascular pigment epithelial detachment as intraretinal fluid.

5.6.2.3.4.5 True positives for CLC and FNs for R6 (n=2)

There were only 2 instances of cases where R6 produced a FN whilst CLC produced a true positive. One appears to be from a VA drop of 7 early treatment of diabetic retinopathy study (ETDRS) letters with stable OCT appearance and the other due to large volumes of IRF being identified on a prior poor-quality scan, numerically concealing the actual trend of increasing IRF between visits (Figure 33).

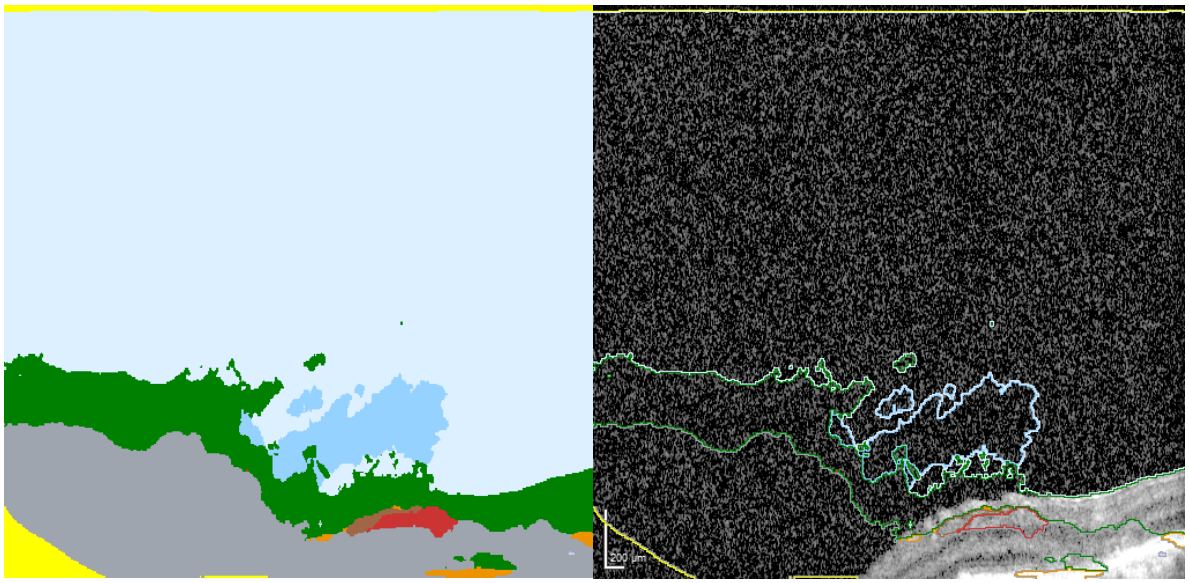


Figure 33. Poor quality B-scan at prior visit leading to a false negative for ruleset 6.

4.6.2.3.4.6 FNs for both CLC and R6 (n=3)

Of these 3 cases it seems 2 were clinically ambiguous. The final one was a data curation error which involved OCTane processing identical OCT scans rather than ones from sequential visits.

5.7 Discussion

The potential for a non-inferior, or even superior safety profile from AI-enabled nAMD treatment decisions compared to CLC is evident. This finding is important because it meets the need that many stakeholders expressed in chapter 4 to see evidence that an AI-enabled intervention would not compromise on current levels of safety and visual outcomes for patients. It also broadens the value proposition of the AI-enabled intervention from pure efficiency gain assumed by stakeholders interviewed in chapter 4 with additional opportunities to improve care quality. This additional value proposition seems at least in part because the quantitative measures of CLC decision safety and accuracy we have derived are lower than would be intuitive for most stakeholders. This was particularly evidenced by the 33 clinic visits (12.6%) where patients had disease activity which was not recognised by CLC.

The relative performance of the AI-enabled decisions depends on which staff groups within CLC they were compared to and which ruleset was applied to derive decisions of nAMD activity. Despite these differences however, the primary outcome of clinically and statistically significant non-inferior rNPV appears to be very robust. This success in avoiding undertreatment looks likely to translate to some increased costs in service provision, as whilst the rPPV of AI-enabled decisions may be close to 1, its greater sensitivity in identifying disease activity means that more intensive treatment intervals would (appropriately) be recommended (Figure 28). This particularly seems to be the case in the loading protocol applied at NuTH over the first year of nAMD treatment. Although this finding cannot evaluate stakeholders' assumptions that AI would improve the efficiency of individual appointments, it seems to suggest that the resource need across a service would increase as a given nAMD population would receive more treatment. Given that each additional treatment requires time contributions from multiple members of staff, patient travel and expensive drug administration even a small increased treatment burden across a service could counteract any efficiency gains that may be made in clinical decision making at the OCT interpretation stage. From the error analysis however, there do appear to be a number of opportunities to improve the rPPV and reduce or reverse these potential additional treatment costs. This could take place through placing a human in the loop to identify gross segmentation errors ("AI interrogation practices") or changing the decision thresholds on SRF to absolute volumes rather than proportional change.[22]

5.7.1 Comparison with prior work

The primary outcome of this work aligns with applied experimental work with OCTane in an adjacent use case to triage referrals for hospital eye service review on the basis of OCT imaging alone. When OCTane categorised OCT images for urgent referral it did so with similar performance to clinicians working off OCT and clinical information.[13] A similar clinician-equivocal performance was demonstrated when an OCTane-based tool was tasked with predicting conversion from AMD to nAMD within 6 months.[23] Notably, both of these prior studies benefited from less disputable reference standards than the present study. For the triage use case, this was a future final diagnosis, whilst for the nAMD prediction use case it was a future observation of nAMD. In the present use case of repeated assessments of disease activity for a known diagnosis of nAMD, such a robust reference standard is unavailable due to the more stochastic way in which disease activity and treatment effect interact over time. It may well be that an 'incorrect' decision at a single clinic visit does not translate to even small amounts of irreversible sight loss for a patient, even though we

know at a population level delayed treatment is associated with greater vision loss in nAMD.[15] Whilst this characteristic of the proposed use case challenges evaluation of the AI in a retrospective setting, it is a great advantage to its implementation, as AI errors in judging disease activity would not often translate into harm.

The nature and frequency of the errors made by OCTane are also consistent with a prior published qualitative assessment of the tool's segmentations.[7] This paper also touched upon the challenging situation where an AI tool is perceived to perform a specialist clinical task to a higher quality than a clinical expert. Clinicians often appear capable of deflecting any professional threats from AI unless it outperforms them on tasks which compose part of their own sense of identity or worth. This is well demonstrated by a qualitative exploration of radiologist and radiographer perspectives of AI-enabled tools.[24] Radiographers appeared to experience a greater threat because the AI threatened a more foundational aspect of their role, whereas radiologists felt that their core value lay elsewhere. From data in chapter 4, it seems that this is mirrored in the present use case as ophthalmologists almost begrudge the amount of their time which is currently spent on what they perceive as the simple and repetitive task of OCT interpretation with an established diagnosis of nAMD. Like the radiographers however, staff with lower positions in the traditional clinical hierarchy do appear to be more wary of the changes that AI may come to impose.[24] Patients may also feel uncomfortable with evidence of AI providing superior clinical decisions for their care. Again, data in chapter 4 demonstrated the esteem in which consultant ophthalmologists are held and AI with apparently superior performance in any regard may feel incongruous and lead patients to hold protective sentiments for their clinicians. The high value which AMD patients assign to ophthalmologist opinion is reported elsewhere as is a cultural scepticism toward the ability of clinical AI from patients and the public. [25-27] This tension was also anticipated by this study's public reference group which spontaneously raised the delicacy of the framing of comparisons between AI and clinician performance when interpreting the results. They recommended a discussion of equivalent performance in narrow tasks and an emphasis on the opportunity for clinician time to be diverted to the more valuable inter-personal elements of care.

5.7.2 Limitations

The retrospective and observational nature of this study poses a significant limitation on the conclusions that can be drawn. However, it is necessary because without such a retrospective evaluation it would be dangerous and wasteful to commit resources to a prospective interventional study and is a recognised step on the translational pathway for clinical AI.[28] The retrospective data do mean that the rationale for CLC decisions is relatively opaque and it is hard to assess whether 'errors', as defined in the study method, should truly be considered as such. For example, a clinician may have verbally acknowledged a patient's aversion for injections and agreed to maintain a treatment interval though they felt the interval should have been reduced. Many stakeholders would consider such shared decision making to represent higher quality care, yet our study method would record such an episode as a sight-threatening FN. During data collection EMR entries were examined for any documented suggestion of shared decision making. Only four such entries were noted across the dataset, the CLC assessment was graded as FP in three of these instances and TN in one. This suggests that shared decision making is uncommonly documented, quite possibly because it uncommonly impacts treatment decisions. However, it does support the limitation described above as it seems these instances are often graded

as ‘errors’ in the retrospective method applied in this chapter. It is also noteworthy that telemedical models of care for nAMD treatment monitoring and the management of other chronic retinal conditions are becoming more prevalent across NHS ophthalmology services, though NuTH operates a face-to-face consultation approach. As such the opportunity for shared decision making in standard care is increasingly not present and so should not necessarily be considered a short coming of the technology itself.

The generalisability of findings from this study are also challenged by their origin from a single site. To justify prospective evaluation in the NHS, replication of the findings will be necessary at other UK sites.

As mentioned in comparisons with prior work, the reference standard of Moorfields Reading Centre judgements is another limitation. It still appears as the closest proxy for a biological ground truth that could be assigned, but by its design it cannot hope to reflect the kind of patient-centred decision making that most clinicians aspire to provide. Even within the confines of its design and aims, the Moorfields Reading Centre process is also subject to human error. However, subsequent presentation and discussion of these data with independent ophthalmologists around the UK has been supportive of the study design, with no suggestions for preferable reference standards advanced in this retrospective setting.

The impact of the present study is also limited by its focus on a non-regulated medical device. Whilst the evidence generated is supportive of any future versions of this device, or analogous regulated devices, it cannot inform any real-world decisions to pursue real-world implementation of such a device. This limitation is a reflection of good practices in clinical AI evidence generation, but is also legally enforceable, due to regulatory frameworks.[29, 30] The use case tested in the present study was also fully automated, rather than the decision support role which is likely to be closer to initial implementation efforts.[31, 32] Whilst this limits the translation of the present study’s findings to potential real-world interventions, it simplifies evaluation of the AI technology itself by avoiding the complex influence that human computer interactions will have on real-world outcomes.[3]

5.7.3 Future directions

The different rule sets and treatment protocols examined in the secondary outcomes have shown that non-inferiority can be achieved over rNPV in a number of different ways with variable trade-offs in the cost, safety and acceptability of the resulting AI-enabled service. Characterising and quantifying these trade-offs whilst exploring others yet to be identified would be a valuable step toward optimising the AI-enabled intervention which would best carry this technology into practice. Some of this is outside of the present scope of work, but a more directed analysis of qualitative data elicited in chapter 4 would be a valuable first step toward crafting an intervention for future prospective evaluation. Such an analysis will form the basis of chapter 6.

Replications of the present non-inferiority study using data from different NHS sites would be another valuable step to take prior to prospective evaluation of the AI-enabled intervention. Ideally this would include clinics that serve varied populations, e.g. ethnicity, use different technical infrastructure, e.g. OCT equipment manufacturers, and draw on different AI-enabled segmentation tools, e.g. a regulated medical device. These replication and health economic aspects are outside of the scope of this thesis.

5.8 Conclusions

This retrospective study provides evidence that AI-enabled nAMD treatment monitoring can be at least as safe as NHS CLC, with clinically insignificant trade-offs in overtreatment for patients. This finding is robust across several different clinical interpretations of OCT segmentation outputs and all the staff groups involved in current nAMD treatment monitoring. It also appears from error analysis that there are further opportunities to improve the performance. These improvements could be realised through adjustments to the AI technology itself, the way the technology's outputs are interpreted within a medical device, the positioning of that device within an AI-enabled intervention or the wider nAMD pathway in which that intervention will sit.

5.9 Appendix

5.9.1 Screening for case characteristics associated with CLC FNs

Table 26. Screening for unequal performance of consultant led care (CLC) between clinic visits with different categorical characteristics using descriptives of the absolute number of false negatives (FN) and the overall false negative rates (FNR) in between different groups. Chi-squared tests are used to derive p values.T2DM = Type 2 Diabetes Mellitus, TEX = Treat-and-Extend.

N=262		not FN	FN	FNR	p value
Number of eyes treated	Unilateral treatment	165	20	12.1%	0.18
	Ongoing bilateral treatment	64	13	20.3%	
Sex	Female	139	22	15.8%	0.51
	Male	90	11	12.2%	
Ethnicity	British	212	31	14.6%	0.91
	Pakistani	1	0	0.0%	
	Not stated	16	2	12.5%	
Laterality	Left	107	15	14.0%	0.89
	Right	122	18	14.8%	
Drug	Aflibercept	207	31	15.0%	0.51
	Ranibizumab	22	2	9.1%	
Diabetic status	Not diabetic	134	20	14.9%	0.86
	T2DM	40	5	12.5%	
	Diabetic - unknown type	4	0	0.0%	
	Status unknown	51	8	15.7%	
Protocol	Loading	44	9	20.5%	0.28
	TEX	185	24	13.0%	

Table 27. Screening for unequal performance of consultant led care (CLC) across clinic visits with different continuous characteristics using descriptives of the mean and 95% confidence intervals (CI) of visits which were and were not assigned a false negative (FN) assessment of disease activity by CLC. Independent t- tests are used to derive p values comparing the difference of the means between visits which were and were not assigned a FN assessment. IDAOPI = Income Deprivation Affecting Older People Index, VA = Visual Acuity, nAMD = neovascular Age-relate Macular Degeneration

N=262 (unless otherwise stated		IDAOPI	Years of age	Baseline VA (letters)	Prior treatment interval	Prior injections	Visit VA (letters)	Contralateral VA (letters)	most recent nAMD activity (weeks) N = 176
not CLC FN n=229	Mean	5.2	80.3	59.6	8.9	12.9	62.8	56.8	26.5
	upper CI	5.6	81.3	61.3	9.3	14.0	64.7	60.5	30.3
	lower CI	4.9	79.4	57.8	8.5	11.8	60.8	53.2	22.8
CLC FN n=33	Mean	5.3	80.5	63.3	8.1	11.6	64.6	57.7	25.8
	upper CI	6.3	83.3	68.5	8.9	14.7	69.8	67.0	33.4
	lower CI	4.4	77.7	58.2	7.3	8.6	59.4	48.3	18.2
	p value	0.81	0.90	0.18	0.08	0.45	0.52	0.87	0.86

5.9.2 Screening for case characteristics associated with CLC FPs

Table 28. Screening for unequal performance of consultant led care (CLC) between clinic visits with different categorical characteristics using descriptives of the absolute number of false positives (FP) and the overall false positive rates (FPR) in between different groups. Chi-squared tests are used to derive p values.T2DM = Type 2 Diabetes Mellitus, TEX = Treat-and-Extend.

N=262		not FP	FP	FPR	p value
Number of eyes treated	Unilateral treatment	150	35	23.3%	0.56
	Ongoing bilateral treatment	60	17	28.3%	
Sex	Female	130	31	23.8%	0.76
	Male	80	21	26.3%	

Ethnicity	British	192	51	26.6%	0.25
	Pakistani	1	0	0.0%	
	Not stated	17	1	5.9%	
Laterality	Left	96	26	27.1%	0.58
	Right	114	26	22.8%	
Drug	Aflibercept	193	45	23.3%	0.23
	Ranibizumab	17	7	41.2%	
Diabetic status	Not diabetic	124	30	24.2%	0.78
	T2DM	36	9	25.0%	
	Diabetic - unknown type	4	0	0.0%	
	Status unknown	46	13	28.3%	
Protocol	Loading	48	5	10.4%	0.03
	TEX	162	47	29.0%	

Table 29. Screening for unequal performance of consultant led care (CLC) across clinic visits with different continuous characteristics using descriptives of the mean and 95% confidence intervals (CI) of visits which were and were not assigned a false positive (FP) assessment of disease activity by CLC. Independent t- tests are used to derive p values comparing the difference of the means between visits which were and were not assigned a FP assessment. IDAOPI = Income Deprivation Affecting Older People Index, VA = Visual Acuity, nAMD = neovascular Age-relate Macular Degeneration

N=262 (unless otherwise stated		IDAOP	Years of age	Baseline VA (letters)	Prior treatment interval	Prior injections	Visit VA (letters)	Contralateral VA (letters)	most recent nAMD activity (weeks) N = 176
not CLC FP n=210	Mean	5.3	80.5	59.7	8.5	12.5	63.1	57.0	26.8
	upper CI	5.7	81.5	61.5	8.8	13.6	65.0	60.8	30.8
	lower CI	4.9	79.5	57.8	8.1	11.3	61.2	53.2	22.7
CLC	Mean	4.9	79.7	61.6	10.0	13.7	62.6	56.8	25.4

FP n=52	upper CI	5.7	81.9	65.6	11.0	16.1	67.2	64.6	31.2
	lower CI	4.1	77.6	57.7	9.0	11.4	58.0	49.1	19.6
	p value	0.36	0.52	0.38	0.01	0.36	0.84	0.98	0.71

5.9.3 Screening for case characteristics associated with R6 FNs

Table 30. Screening for unequal performance of OCTane outputs interpreted by rule set 6 (R6) between clinic visits with different categorical characteristics using descriptives of the absolute number of false negatives (FN) and the overall false negative rates (FNR) in between different groups. Chi-squared tests are used to derive p values.T2DM = Type 2 Diabetes Mellitus, TEX = Treat-and-Extend.

N=262		not FN	FN	FNR	p value
Number of eyes treated	Unilateral treatment	182	3	1.6%	0.60
	Ongoing bilateral treatment	75	2	2.7%	
Sex	Female	157	4	2.5%	0.39
	Male	100	1	1.0%	
Ethnicity	British	238	5	2.1%	0.82
	Pakistani	1	0	0.0%	
	Not stated	18	0	0.0%	
Laterality	Left	118	4	3.4%	0.13
	Right	139	1	0.7%	
Drug	Aflibercept	234	4	1.7%	0.40
	Ranibizumab	23	1	4.3%	
Diabetic status	Not diabetic	150	4	2.7%	0.65
	T2DM	44	1	2.3%	
	Diabetic - unknown type	4	0	0.0%	
	Status unknown	59	0	0.0%	
Protocol	Loading	53	0	0.0%	0.26
	TEX	204	5	2.5%	

Table 31. Screening for unequal performance of OCTane outputs interpreted by rule set 6 (R6) across clinic visits with different continuous characteristics using descriptives of the mean and 95% confidence intervals (CI) of visits which were and were not assigned a false negative (FN) assessment of disease activity by CLC. Independent t- tests are used to derive p values comparing the difference of the means between visits which were and were not assigned a FN assessment. IDAOPI = Income Deprivation Affecting Older People Index, VA = Visual Acuity, nAMD = neovascular Age-relate Macular Degeneration

N=262 (unless otherwise stated		IDAOP	Years of age	Baseline VA (letters)	Prior treatment interval	Prior injections	Visit VA (letters)	Contralateral VA (letters)	most recent nAMD activity (weeks) N = 176
not R6 FN n=257	Mean	5.2	80.4	60.2	8.7	12.7	63.0	56.8	26.0
	upper CI	5.6	81.3	61.9	9.1	13.7	64.9	60.3	29.4
	lower CI	4.9	79.4	58.6	8.4	11.6	61.2	53.4	22.7
R6 FN n=5	Mean	4.8	81.0	51.4	10.0	15.6	61.2	63.4	40.6
	upper CI	6.4	85.1	62.7	17.5	22.3	74.5	87.1	73.5
	lower CI	3.2	76.9	40.1	2.5	8.9	47.9	39.7	7.7
	p value	0.62	0.78	0.20	0.76	0.44	0.80	0.62	0.44

5.9.4 Screening for case characteristics associated with R6 FPs

Table 32. Screening for unequal performance of OCTane outputs interpreted by rule set 6 (R6) between clinic visits with different categorical characteristics using descriptives of the absolute number of false positives (FP) and the overall false negative rates (FPR) in between different groups. Chi-squared tests are used to derive p values.T2DM = Type 2 Diabetes Mellitus, TEX = Treat-and-Extend.

N=262		not FP	FP	FPR	p value
Number of eyes treated	Unilateral treatment	124	61	49.2%	0.47
	Ongoing bilateral treatment	48	29	60.4%	
Sex	Female	107	54	50.5%	0.73
	Male	65	36	55.4%	
Ethnicity	British	160	83	51.9%	0.38
	Pakistani	0	1	100.0%	

	Not stated	12	6	50.0%	
Laterality	Left	82	40	48.8%	0.25
	Right	90	50	55.6%	
Drug	Aflibercept	159	79	49.7%	0.21
	Ranibizumab	13	11	84.6%	
Diabetic status	Not diabetic	106	48	45.3%	0.59
	T2DM	28	17	60.7%	
	Diabetic - unknown type	2	2	100.0%	
	Status unknown	36	23	63.9%	
Protocol	Loading	32	21	65.6%	0.82
	TEX	140	69	49.3%	

Table 33. Screening for unequal performance of OCTane outputs interpreted by rule set 6 (R6) across clinic visits with different continuous characteristics using descriptives of the mean and 95% confidence intervals (CI) of visits which were and were not assigned a false positive (FP) assessment of disease activity by CLC. Independent t- tests are used to derive p values comparing the difference of the means between visits which were and were not assigned a FN assessment. IDAOPI = Income Deprivation Affecting Older People Index, VA = Visual Acuity, nAMD = neovascular Age-relate Macular Degeneration

N=262 (unless otherwise stated		IDAOP	Years of age	Baseline VA (letters)	Prior treatment interval	Prior injections	Visit VA (letters)	Contralateral VA (letters)	most recent nAMD activity (weeks) N = 176
not R6 FP n=172	Mean	5.3	80.7	60.9	8.7	13.2	63.6	57.6	24.1
	upper CI	5.7	81.8	62.9	9.1	14.6	65.8	61.8	27.8
	lower CI	4.9	79.6	58.9	8.3	11.9	61.5	53.3	20.3
R6 FP n=90	Mean	5.1	79.7	58.4	9.0	11.8	61.8	55.8	32.1
	upper CI	5.7	81.3	61.3	9.6	13.4	65.0	61.5	39.1

	lower CI	4.5	78.1	55.5	8.3	10.2	58.6	50.1	25.2
	p value	0.71	0.31	0.16	0.44	0.18	0.35	0.63	0.05

5.10 References

1. The United States Food and Drug Administration, FDA Approved Drugs: New Drug Application (NDA): 217171. 2023.

2. The Medicines and Healthcare products Regulatory Agency, MHRA announces new recognition routes to facilitate safe access to new medicines with seven international partners. 2023.

3. Cabitza, F., et al., Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. Artificial Intelligence in Medicine, 2023. 138: p. 102506.

4. Hogg, H.D.J., et al., Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence. J Med Internet Res, 2023. 25: p. e39742.

5. Suresh, H. and J. Guttag, A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, in Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. 2021, Association for Computing Machinery: --, NY, USA. p. Article 17.

6. Software as a Medical Device Working Group, Software as a Medical Device (SaMD): Clinical Evaluation - IMDRF/SaMD WG/N41FINAL:2017, I.M.D.R. Forum, Editor. 2017.

7. Wilson, M., et al., Validation and Clinical Applicability of Whole-Volume Automated Segmentation of Optical Coherence Tomography in Retinal Disease Using Deep Learning. JAMA Ophthalmology, 2021. 139(9): p. 964-973.

8. Russell-Puleri, S., et al., Comparison of a Deep Learning based OCT image segmentation algorithm to manual segmentation by a traditional reading center for patients with wet AMD. Investigative Ophthalmology & Visual Science, 2023. 64(8): p. 316-316.

9. Ross, A.H., et al., Recommendations by a UK expert panel on an aflibercept treat-and-extend pathway for the treatment of neovascular age-related macular degeneration. Eye, 2020. 34(10): p. 1825-1834.

10. Moorfields Eye Hospital NHS Foundation Trust, Moorfields Eye Hospital Ophthalmic Reading Centre and Clinical AI Lab. 2023; Available from: https://readingcentre.org/.

11. Walker, E. and A.S. Nowacki, Understanding equivalence and noninferiority testing. J Gen Intern Med, 2011. 26(2): p. 192-6.

12. Annastazia, E.L., et al., FENETRE study: quality-assured follow-up of quiescent neovascular age-related macular degeneration by non-medical practitioners: study protocol and statistical analysis plan for a randomised controlled trial. BMJ Open, 2021. 11(5): p. e049411.

13. De Fauw, J., et al., Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med, 2018. 24(9): p. 1342-1350.

14. Hogg, J. The prevalence and impact of treatment delays in exudative age-related macular degeneration. 2021. Investigative Ophthalmology & Visual Science.

15. Fu, D.J., et al., Insights From Survival Analyses During 12 Years of Anti–Vascular Endothelial Growth Factor Therapy for Neovascular Age-Related Macular Degeneration. JAMA Ophthalmology, 2021. 139(1): p. 57-67.

16. Ji Eun Diana, H., et al., Teleophthalmology-enabled and artificial intelligence-ready referral pathway for community optometry referrals of retinal disease (HERMES): a Cluster Randomised Superiority Trial with a linked Diagnostic Accuracy Study—HERMES study report 1—study protocol. BMJ Open, 2022. 12(2): p. e055845.

17. Moskowitz, C.S. and M.S. Pepe, Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clin Trials, 2006. 3(3): p. 272-9.

18. Teare, M.D., et al., Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. Trials, 2014. 15(1): p. 264.

19. Barnaby, C.R., et al., Effectiveness of Community versus Hospital Eye Service follow-up for patients with neovascular age-related macular degeneration with quiescent disease (ECHoES): a virtual non-inferiority trial. BMJ Open, 2016. 6(7): p. e010685.

20. CLOPPER, C.J. and E.S. PEARSON, THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL. Biometrika, 1934. 26(4): p. 404-413.

21. Andersson, P.G., The Wald Confidence Interval for a Binomial p as an Illuminating “Bad” Example. The American Statistician, 2023. 77(4): p. 443-448.

22. Lebovitz, S., et al, To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. Organization Science, 2022. 33(1): p. 126-148.

23. Yim, J., et al., Predicting conversion to wet age-related macular degeneration using deep learning. Nature Medicine, 2020. 26(6): p. 892-899.

24. Chen, Y., et al., Professionals’ responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. BMC Health Services Research, 2021. 21(1): p. 813.

25. Boyle, J., et al., Experiences of patients undergoing anti-VEGF treatment for neovascular age-related macular degeneration: a systematic review. Psychol Health Med, 2015. 20(3): p. 296-310.

26. Jelin, E., et al., Development and testing of a patient-derived questionnaire for treatment of neovascular age-related macular degeneration: dimensions of importance in treatment of neovascular age-related macular degeneration. Acta Ophthalmol, 2018. 96(8): p. 804-811.

27. The Accelerated Access Collaborative, Public perceptions and attitudes to Artificial Intelligence (AI) in healthcare. An exploratory study. 2022.

28. Vasey, B., et al., Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 2022. 28(5): p. 924-933.

29. Liu, X., et al., Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med, 2020. 26(9): p. 1364-1374.

30. The Medicines and Healthcare products Regulatory Agency, Crafting an intended purpose in the context of Software as a Medical Device (SaMD). 2023.

31. Holz, F.G., et al., Does real-time artificial intelligence-based visual pathology enhancement of three-dimensional optical coherence tomography scans optimise treatment decision in patients with nAMD? Rationale and design of the RAZORBILL study. Br J Ophthalmol, 2023. 107(1): p. 96-101.

32. Coulibaly, L.M., et al., Personalized treatment supported by automated quantitative fluid analysis in active neovascular age-related macular degeneration (nAMD)—a phase III, prospective, multicentre, randomized study: design and methods. Eye, 2023. 37(7): p. 1464-1469.

Chapter 6: Proposing a specific AI-enabled intervention for nAMD treatment monitoring

Problem: Prior chapters have outlined an AI-enabled device for nAMD treatment monitoring which appears acceptable to stakeholders. To integrate with patient care, such a device must be carried within a healthcare intervention, defining who will interact with the device, how and when. To promote successful implementation, the various facets of such an intervention should be designed to align with the determinants of implementation surfaced in chapter 3. However, the evaluative framework used to produce these determinants does not readily derive an intervention and so a secondary, more goal-focused analysis is required.

Objectives: This secondary analysis of the data and implementation determinants derived in chapter 3 aims to recommend an evidence-based AI-enabled intervention for nAMD treatment monitoring, optimised for implementation success.

Methods: Drawing on the TMFs used in qualitative clinical AI research curated in chapter 3, a tripartite process model, the Fit between *Individual*, *Technology* and *Task* (FITT) Framework, was identified to support intervention design. The intervention underwent formative iterative review by patient, public and multidisciplinary professional members of the Study Reference and Advisory Groups.

Findings: Considering the *task* domain, nAMD treatment should be initiated at F2F appointments with clinicians who recommend year-long periods of autonomous AI-enabled scheduling of treatments. Lines of communication for emerging patient concerns about symptoms or treatment plans will be maintained and at least one high-quality F2F consultation will take place annually. Considering the *individual* domain, appropriately trained photographers should take on the additional roles of inputting retinal imaging into the AI device and overseeing its communication to clinical and administrative colleagues. Ophthalmologists would be responsible for clinical oversight, annual F2F consultations and handling patient or photographer concerns arising across the year of AI-enabled treatment monitoring. Considering the *technology* domain, interoperability to facilitate this intervention would best be served by imaging equipment that can send images to the cloud securely for analysis by AI tools. Picture Archiving and Communication Software (PACS) should have the capability to output directly into the EMR used by clinical and administrative staff.

Conclusions: This secondary theory-informed analysis has proposed an AI-enabled intervention which can facilitate prospective evaluation and consideration of a full hypothetical AI-enabled care pathway.

Relevance to future chapters: The proposed AI-enabled intervention for nAMD treatment monitoring forms the basis for a formal Intended Use Statement, which will ultimately be required for regulatory approval and clinical use. The additional level of detail also facilitates the early identification and mitigation of risks that may emerge indirectly from the AI technology at distant points in the healthcare pathway. A medical algorithmic audit will be performed in chapter 7 to deliver both elements.

6.1 Background

A health technology intervention is a sociotechnical vehicle for the implementation of health technologies such as AI. These interventions hold various specifications about how, when and where the technology should be used, by who, for who and for what purpose.[1] There is a great range of potential detail that can be specified for each of these aspects and anything that is not specified in advance must be determined by the individuals interacting with the technology.[2] To minimise the variability in health technology usage, and the risks associated with such variability, regulators incentivise manufacturers to set objective intended uses for SaMD.[1] These requirements mean that even identical technologies are considered as distinct SaMD when their intended use is changed. The associated need for renewed regulatory approvals conveys significant disincentives to vendors who seek to dramatically alter the intervention after they have achieved an initial regulatory approval. Other key decision makers such as Health Technology Assessment bodies, commissioners and healthcare managers also exert their own overlapping disincentives on shifting the scope of health technology interventions that they have endorsed.

Despite these regulatory and contractual restrictions, the design of an AI-enabled healthcare intervention is one of the most modifiable yet influential determinants of implementation success.[3] This is demonstrated by the NASSS framework itself, but also cross-specialty clinical AI research (Chapter 3) and the qualitative data elicited from stakeholders in potential AI-enabled nAMD treatment monitoring (Chapter 4).[4] A qualitative study of US academic medical centres using clinical AI, also illustrated a strong preference from leaders in AI implementation to align AI-enabled interventions with current ways of working, rather than vice versa.[5] When returning to the residual pragmatic question from chapter 4 of *how* AI-enabled nAMD treatment monitoring should be enacted, intervention design is therefore fundamental.

One of the most mature AI-enabled interventions in monitoring retinal disease is for diabetic retinopathy grading, with many live AI-enabled healthcare pathways internationally.[6] There are more than 20 separate regulated AI medical devices that could be used for such an intervention in various jurisdictions across the globe.[7] Some investigators have considered the implications of these differing implementation contexts when designing the intervention in which their AI technology sits. This is simply illustrated by the extremely different thresholds of sensitivity and specificity selected by the same manufacturer for two pivotal trials which target distinct contexts in the US and UK.[8, 9] The UK has had a high-performing diabetic retinopathy screening service for many years and so any potentially valuable intervention must demonstrate extremely high sensitivity (missing an extremely low proportion of disease positive individuals). This aspect of implementation context explains the selection of a diagnostic threshold with 100% (95% CI 98.7%, 100%) sensitivity and 54% (95% CI 53.4%, 54.5%) specificity for the UK-based trial.[8] Contrastingly, the US has a more fragmented and less comprehensive service with a higher cost to payors and patients from unnecessary referrals. As such, the value proposition is not so dependent on high sensitivity, permitting greater prioritisation of specificity. This is demonstrated by the threshold with 95.5% (95% CI 92.4%, 98.5%) sensitivity and 85.0% (95% CI 82.6%, 87.4%) specificity selected for the same technology in the US-based trial.[9] Even with these considerations and substantial investment by healthcare decision makers, the high clinical and cost-effectiveness requirements of established NHS systems have prevented the

implementation of AI-enabled diabetic retinopathy screening in England (limited use in Scotland).[10]

Guidelines for the design and evaluation of complex healthcare interventions, such as those involving AI, are already established.[11] The methods applied within the disciplinary field of implementation science address many of these guidelines. In particular, the value of carefully selected TMFs in harnessing data for evidence-based intervention evaluation and design.[12, 13] Despite broad consensus on the importance of TMF-informed intervention design and evaluation, which TMFs to use and how to use them remain as decisions to be made on a case-by-case basis by researchers. There are a great number of TMFs available to choose from with great variation in their selection, even within the narrow confines of clinical AI research (chapter 3).[14]

6.2 Problem

Chapter 4 identified what is likely to influence the implementation of AI-enabled nAMD treatment monitoring and why. It also identified what stakeholders would perceive as successful change in nAMD care. What remains unclear is how the design of an AI-enabled intervention can best operationalise these insights into implementation determinants and success criteria to deliver value to stakeholders in nAMD care.

Failure of healthcare interventions is the norm.[15] If the absence of an evidence-based AI-enabled intervention for nAMD treatment monitoring persists, any potential implementation efforts seem unlikely to escape this expectation of delay or failure .[16, 17] Any such failures may damage stakeholders' perspective of the health technology, further lowering the long-term chances of implementation.[18]

The determinants of implementation success vary over time and between target contexts.[4] If it is to remain effective, any evidence-based intervention should be adaptable to these changes and so the evidence and rationale that guided prior iterations of the intervention needs to be clear and transparent.

6.3 Rationale

To optimise the chance of implementation success for AI-enabled nAMD treatment monitoring, this chapter adopts best practices from implementation science (Figure 18).[12] Stakeholder groups to clinical AI implementation have been rigorously identified (Chapter 3), their perspectives on AI-enabled nAMD treatment monitoring have been analysed to identify determinants of implementation (Chapter 4) and hypothetical patient outcomes have been investigated (chapter 5). Now, a carefully selected TMF will be used to harness these insights to shape the intervention in which the AI technology should sit. Because of the practical goal to determine how AI-enabled nAMD treatment monitoring should be implemented, the TMF should be an action-oriented process model.[19] It should expose the sociotechnical mechanisms by which various elements of an intervention are expected to influence its implementation. The TMF should also only concern domains which are modifiable through intervention design, helping to focus the data on the task at hand and maintain meaningful transparency to varied stakeholders of the intervention.

The rationale in completing and disseminating this work prior to NHS adoption of an AIaMD for nAMD treatment monitoring, is to inform and improve the design of future relevant AIaMD from all potential manufacturers. This should at least partially mitigate the

disconnect between different stakeholder groups in clinical AI design and development, which has so often limited implementation efforts elsewhere in the literature.[3] The style of writing in the results section is consciously distinct from that of chapter 4 to make the evidence more accessible, and therefore useful, to the practice-orientated decision makers for who it is intended (chapter 2).[20] The TMF will also act as a record of the evidence and rationale that guided the proposed intervention, permitting later re-iterations as the sociotechnical context which it targets evolves.[4]

6.4 Aim

This secondary analysis of data and findings from chapter 3 aims to begin to answer how AI-enabled nAMD treatment monitoring should be implemented. Due to its relatively high modifiability and influence, this is achieved through a focus on the design of the AI-enabled intervention that should carry the technology into practice.

6.5 Methods

All TMFs identified through the systematic search of primary qualitative research into clinical AI were reviewed for relevance to this chapter's aim.[14] The Fit between Individuals, Technology and Task Framework was identified as a relatively simplistic, action-oriented model, focused upon elements of an intervention which are modifiable.[21] Two leading alternatives were Davis' TAM and Sittig and Singh's Sociotechnical model.[22, 23]

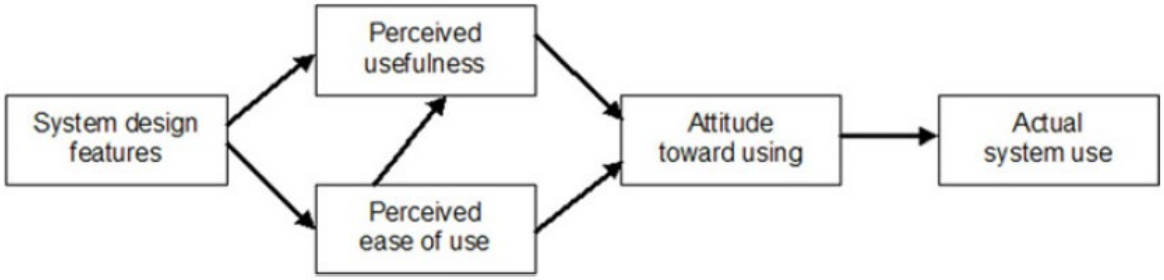


Figure 34. Technology Acceptance Model schematic.[22]

Both are well suited to facilitating thoughtful intervention design as they promote consideration of different facets of an intervention and highlight consequences of those design choices that are likely to influence implementation.[24, 25] However, TAM (Figure 34) focuses more heavily on individual attributes such as the behavioural and psychological responses to interventions by adopters, which can only be indirectly influenced by intervention design choices. Meanwhile, the Sociotechnical Model (Figure 35) highlights numerous and diverse aspects of intervention characteristics and their implementation context. This is helpful in more fully understanding the aspects of an intervention that may influence its implementation but risks distraction with non-modifiable aspects of an intervention or the target context.

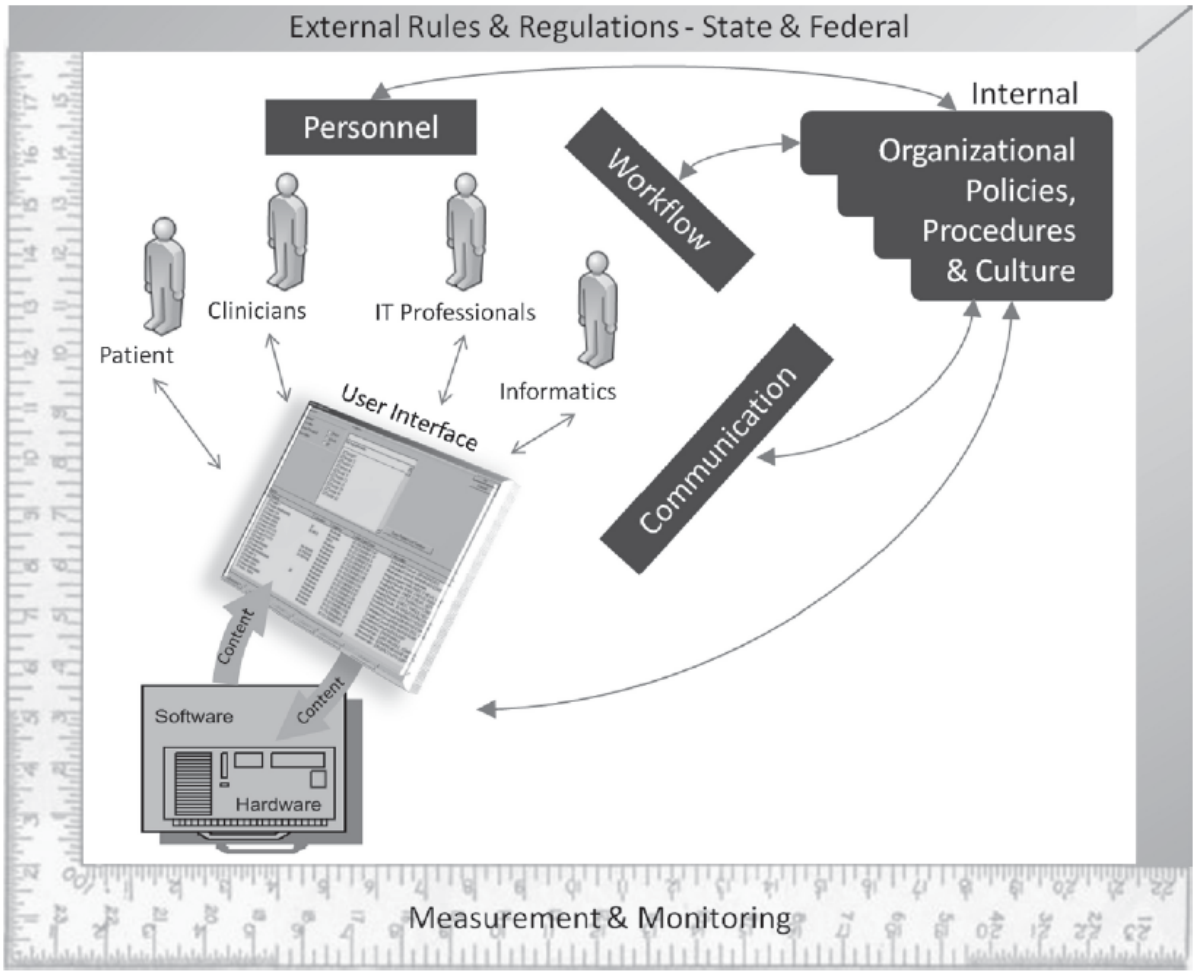


Figure 35. Schematic of Sittig and Singh's Sociotechnical Model.[23]

Following re-familiarisation with the primary data and analysis from chapter 4, each of the three domains of Individuals, Task and Technology were then reviewed (Figure 36). Determinants which appeared related to one of the domains were used to shape an aspect of the intervention, with notes made to document the rationale of that design consideration (see appendix 6). When determinants relevant to a single aspect of the intervention were poorly aligned, an initial judgement was made by the lead researcher to compromise or prioritise between determinants. This process was completed independently by the lead researcher until a draft intervention was complete. This draft was then discussed with the study Reference and Advisory Groups (see appendix 4) to gain additional perspectives on intervention design and to support further iterations of the intervention. The aspects of the intervention which were based on conflicting data were made a particular focus for these discussions. Following this analysis, the proposed intervention was then presented and validated through parallel roundtable discussions at a F2F public engagement event in Newcastle on 4th October 2023. This event was attended by 34 members of the public, ophthalmology patients, eye charity professionals, academics and clinicians (see appendix).

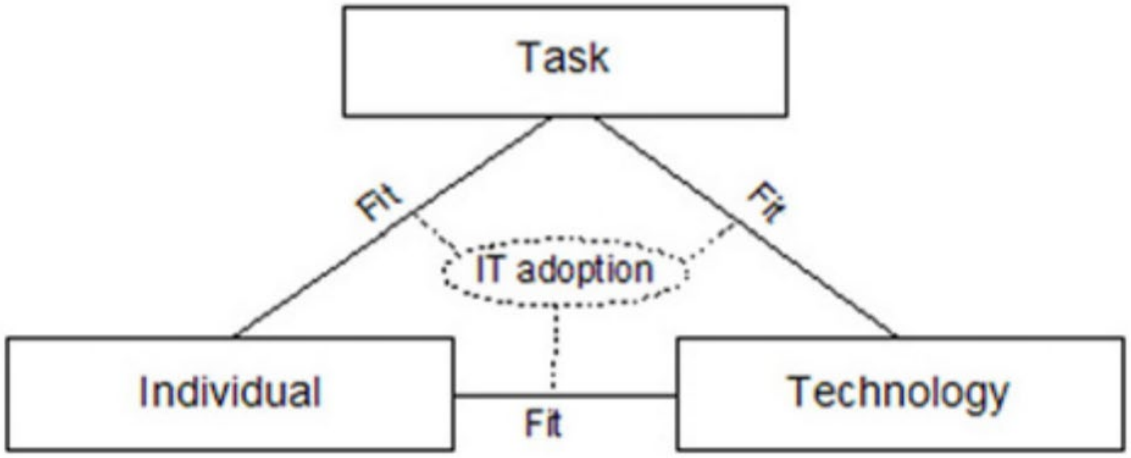


Figure 36. The Fit between Individual, Technology and Task (FITT) framework.[21]

6.6 Results

6.6.1 FITT framework – Task

The data analysis indicates that the AI technology should focus on treatment monitoring for patients with an established diagnosis of nAMD and should incrementally assume relatively autonomous decision making roles (rather than purely supportive) for at least clinical decisions.

Points in the care pathway that deal with clinical and interpersonal complexity such as diagnosis, screening for ocular comorbidities and alterations to management plans were seen to require specialist clinician in input. One patient emphasised that he "*wouldn't like to go straight from AI diagnosis to treatment*" but did value the prospect of extra clinician time being focused on diagnostic consultations, "*I mean, no one spoke about this. They just said, oh, you've had a [macula] haemorrhage*". AI should instead target the high volume and low complexity situations in nAMD care which one ophthalmology trainee described as "*when we are comfortable with the diagnosis and they are on a treatment*" and "*we just need something that can crunch [data]*". Due to existing treatment protocols, the risk of patients losing vision from automating a proportion of such decisions was seen as low. With adequate signposting and safety netting from clinicians, patients appeared very accepting of dropping consultations from most of their treatment appointments, "*that wouldn't bother me if they thought I didn't need to see them all the time"* [Patient]. The firm upper limit for periods of automated monitoring was not explored but no participants was supportive of periods of automated monitoring longer than 12 months. There was also a clear expectation from clinicians and patients that these residual face-to-face appointments would facilitate higher quality discussions and more holistic clinical considerations, e.g. checking for ocular co-morbidities, visual driving standard compliance and sight loss registration.

Setting the AI this relatively independent task would help to clarify the productivity value proposition of AI-enabled treatment monitoring. Making such a clear value proposition would strongly incentivise payors, with one commissioner explaining that "*any innovation that would reduce your backlog by 50%, by 25%, they would just commission it*". Given the mean 6.98 injections per year observed for nAMD patients in chapter 5, most of which were accompanied by distinct consultations, delivering these productivity gains would appear

readily achievable with even minimal frequencies of decision automation. Whilst this level of automation should be the end-goal for an AI-enabled intervention, it is important that it is reached iteratively in partnership with patients, clinicians and managers. All patients should feel able to access evidence of AI safety, though many will not feel the need to see that evidence if their clinical team appears supportive. One patient with a scientific background said they would want to see that AI decisions on their own prior imaging agreed with their clinician's and only then would permit autonomous AI decisions "*for the next two or three [treatments] and then I will come down and*" have another consultation. Clinicians echoed this need for personal evaluation of technology and a gradual trust-building process. A manager reflected on the positive influence of a local evaluation on subsequent treatment innovation, anticipating that both for patients and staff, such a transparent evaluation of AI-enabled nAMD monitoring would, "*give you that sense of pride and ownership*".

5.6.2 FITT framework – Individuals

The data analysis suggests that consultant ophthalmologists should remain accountable and contactable, whilst reducing their contribution to conducting injections and consultations. A larger pool of nursing staff should be recruited to less intensive injection duties. Medical photographers should be responsible for operational aspects of AI use.

An AI-enabled intervention should enable managers' broader goals "*to diversify our workforce, so we have lower qualified and non-qualified staff*" [NHS hospital directorate manager] to take monitoring tasks away from scarce and expensive ophthalmologists (Table 34). A GP warned this should not compromise patients' sense of connection with consultants as their patients already report low satisfaction from seeing trainees, *"if it's, "I was only seen by the algorithm," I can imagine that being an even lower drop*" [GP]. Ideally this would be addressed by prioritising consultants' time for initial diagnostic appointments as patients enter the service and through their oversight of asynchronous lines of communication, laid on for patients' concerns or questions. The direct management of these communications should be delegated across clinical and administrative colleagues. The relative auditability of AI seemed to help consultants to accept this more distant role with one retina specialist consultant reflecting that they are already accountable for the decisions of junior colleagues though he had *"not actually measured how well they perform"*. The intervention should also free up highly qualified allied health professionals and trainee ophthalmologists to spend more time communicating with patients and using their wider competencies. "*I don't see it as a good use of my time after my 10 years of training to look at an OCT for someone who's already got a diagnosis.... I mean, why*?" [ophthalmology trainee]. Freeing up clinicians would deliver value across the ophthalmology service, with one senior manager remarking that whilst macula services were a pain point, its strong leadership meant they "*don't lose sleep over it like [they] do in other areas of ophthalmology*" [NHS hospital directorate manager].

Interoperability issues within and between the various health technologies involved require a staff group to take responsibility for applying the AI and documenting its outputs. Within the hospital setting, hospital photographers represent an accepted and accepting group for this task. One ophthalmic photographer shared the favourite part of their current role was "*when you help diagnose something*" whilst a separate ophthalmic photography lead felt his staff would be very happy to take responsibility for AI on as "*it fixes you very firmly in a critical role"*. When asked who they thought should take responsibility for the AI one patient

answered, "*I think it should be the people that do the photographs now*". Whilst most patients wanted overall responsibility for the AI to be held by a consultant ophthalmologist, none required that relaying or explaining AI outputs involved ophthalmologists and many preferred it not to.

Having a member of the team dedicated to absorbing the technical complexity of AI integration also maximises the productivity of clinicians focused on delivering injections. This is because for injection staff, "*I'm not in the role of a [Band] 7 where I'm reviewing patients. I'm interested in what the diagnosis is, bang, bang, bang, what, what drug, what eye, when to bring them back"* [Advanced nurse practitioner]. The AI output that photographers add to the electronic record should reflect these practical requirements from injectors should also take account of other stakeholders needs as concisely as possible (see appendix). One injector anticipated that with standardised AI documentation they "*wouldn't need to go and ask for as much help*" [Advanced nurse practitioner], which involves physically finding a colleague elsewhere in the department to clarify ambiguous documentation, also disrupting their workflow. Whilst injection sessions were welcomed by nurse and optometrist participants in moderation, it was felt important to keep a good "*balance of the injections so you're not all stuck on injections all day*" [Hospital optometrist]. Along with considerations for clinician satisfaction, taking large amounts of time of band 7 staff to perform duties within the scope of band 6 staff is not cost effective (Table 34). The clinic manager shared a prior solution to this where "*all of the Band 5 nurses, they had the opportunity to inject in the department. And they would be paid as a Band 6 sessional*" [Clinic charge nurse]. This means the band 5 nurses take on better paid work, increase the variety in their job plan, gain exposure to greater clinical responsibility and improve their prospects of career progression. Meanwhile, the relatively small pool of band 7 nurses does not find themselves burdened with a job plan consisting mainly of injection clinics and make use of their examination and consultation skills.

Table 34. Professional groups, roles and lowest associated pay scale for Artificial Intelligence (AI)-enabled neovascular age-related macular degeneration (nAMD) monitoring intervention. AFC = Agenda for Change, OCT = Optical Coherence Tomography, ST = Specialty Training year, F2F = Face-to-Face. [26, 27]

Task	Individual	Pay scale	Starting salary
Checking patients in/out and booking appointments	Receptionist	AFC – band 2	£22,383
OCT capture and AI input/output	Medical photographer	AFC – band 5	£28,407
Injection delivery	Band 5 clinic staff nurses	AFC – pro rata band 6 rate	£35,392
Injection assistance	Healthcare assistants	AFC – band 3	£22,816
Visual acuity measurement	Healthcare assistants	AFC – band 3	£22,816

F2F consultations and managing additional patient messages	Advanced nurse practitioners/Hospital optometrists	AFC – band 7	£43,742
F2F consultations and managing additional patient messages	Junior ophthalmologists	Junior doctor ST1-2	£43,923
Monitoring service performance and providing advice to colleagues	Consultant ophthalmologists	Consultant	£93,666

5.6.3 FITT framework – Technology

The data suggest that the commercial risk of trying to lead competitors in product development mean that initial AIaMD products are likely to come from a small commercial vendor. The AIaMD will need to be compatible with the cloud platform, Picture Archiving and Communication System (PACS), EMR and imaging file formats in the adopter's established digital infrastructure. AIaMD outputs will ideally be smoothly integrated into clinical, administrative and other professionals' workflows.

Although AI for nAMD decision support from OCT analysis could be supplied to providers as a stand-alone software to be integrated onsite, *"there will be a lot of work on the hospital side to do that integration"* [Medtech industry professional]. Such an approach with AI was also described as risky as subtle changes in input data "*could impact the result in ways that are unintended, so the software needs constant monitoring"* [Medtech industry professional]. Industry participants were sceptical of most NHS organisations' capacity or capability to undertake that kind of integration and monitoring work. Consequently, the AIaMD should access provider data on the cloud, as *"having a cloud-based solution lets the vendor be involved in the ongoing surveillance of the performance of their product"* [Medtech industry professional]. For large MedTech companies, cloud-based platforms and other digital infrastructures are one of the secure requirements for AI-enabled healthcare, and so this perception of low commercial risk had motivated significant resource has been committed to developing them. These larger companies appear to be avoiding developing the AIaMD themselves, because it is a risker investment and because such a company "*doesn't want to be burdened with all the regulations and all the issues of developing and clearing such products"* [Medtech industry professional]. This means that the interoperability issues between imaging format, PACS, cloud platform and AIaMD will have been addressed by the larger established companies. This is particularly positive for scalability as like its competitors, the ophthalmic imaging company we spoke to "*have a really great installed base of diagnostic imaging systems and PAC systems and EMR in the case of the UK"* (Table 35).

Table 35. Technological components of the artificial intelligence (AI)-enabled intervention, with descriptions of their role and the proposed vendor for Newcastle upon Tyne Hospitals NHS Foundation Trust (NuTH). OCT = Optical Coherence

Tomography PACS = Picture Archiving and Communication System), nAMD = neovascular age-related macular degeneration, EMR = Electronic Medical Record

Technological component	Purpose	Proposed vendor and rational for NuTH	Anticipated additional cost
OCT equipment	To capture OCT imaging for clinician and AI review	Heidelberg Engineering™; NuTH currently exclusively owns Heidelberg OCT equipment and their outputs are compatible with other relevant technologies	None
PACS	To access, store, send and receive OCT imaging for clinicians and other information systems	Heidelberg Engineering™; NuTH currently uses a Heidelberg PACS (Heyex 2.6) and it is compatible with other relevant technologies	None
EMR	To access, store, send and receive clinical and administrative information for clinicians and other information systems	No clear rationale; NuTH has recently procured Medisite™ which is owned by Heidelberg Engineering™	None
Cloud-based platform	To facilitate OCT imaging input from PACS for AI medical devices and output to EMR	Appway™; As a Heidelberg product interoperability with a version of their PACS (Heyex 2.6.3) is assured and they have already contract two separate OCT segmentation AI products	None
AI medical device	To analyse OCT imaging and output interval recommendation for next nAMD treatment	No clear rationale; RetinAI™ and RetinSight™ are both companies with regulated AI products for OCT segmentation contracted to Appway™, but neither currently offer treatment interval recommendation	Closest products currently charge €5 (£4.30) per application with no current precedent for a NHS customer

The AIaMD remains a foundational part of the intervention however, and the manufacturer "*needs to be a legally registered entity and take legal responsibility for the ownership, distribution, sale and eventual killing of the device*" [Regulatory professional]. The regulatory participant participating in the study was not aware of any NHS organisation acting as

manufacturer for AIaMD to date. Industry participants pointed to a handful of small companies who they are collaborating with to supply AIaMD to their cloud platform. The nature of this arrangement adds some complexity "*where the site that uses it would have to sign a contract"* [Medtech industry professional] in addition to its procurement of the cloud platform from the larger MedTech company. The unfamiliarity of these companies is also complicated by all of the ones mentioned being based outside of the UK, potentially necessitating the egress of sensitive data abroad, which could cause concerns over data governance. At least some mitigations to this as the vendor of one such cloud platform explained that as the data is egressed it "*is completely anonymized and stripped away of any patient information"* [Medtech industry professional].

Ideally, the digital infrastructure which the AIaMD interacts with should allow smooth integration of its outputs with the workflow of different professional groups. Clinicians want some sense of "*how happy it is with the answer it's given you*" and "*would definitely prefer to know exactly why the machine comes up with*" [ophthalmology trainee] a given recommendation. Integration with appointment scheduling facets of the EMR, which have not been interoperable with clinical components of the EMR, should also be pursued. One NHS administrative manager thought that an AI-enabled workflow could improve the safety and efficiency of the service's work as they thought "*it could do away with a lot of having to monitor the appointments"* and anticipated more accurate forecasts and better use of clinical capacity. There may also be opportunities to improve more holistic aspects of nAMD care if AIaMD outputs could be integrated with the social service liaison team. A representative reported significant opportunities to make sure "*referrals are done appropriately and at a relevant time and that the people who do need support are getting the opportunity to access it*" [Social care liaison officer]. Alerts to motivate these referrals and other more holistic considerations (e.g. certification of visual impairment) could be embedded within an AI-enabled workflow to improve, rather than inhibit, the human-touch of nAMD services.

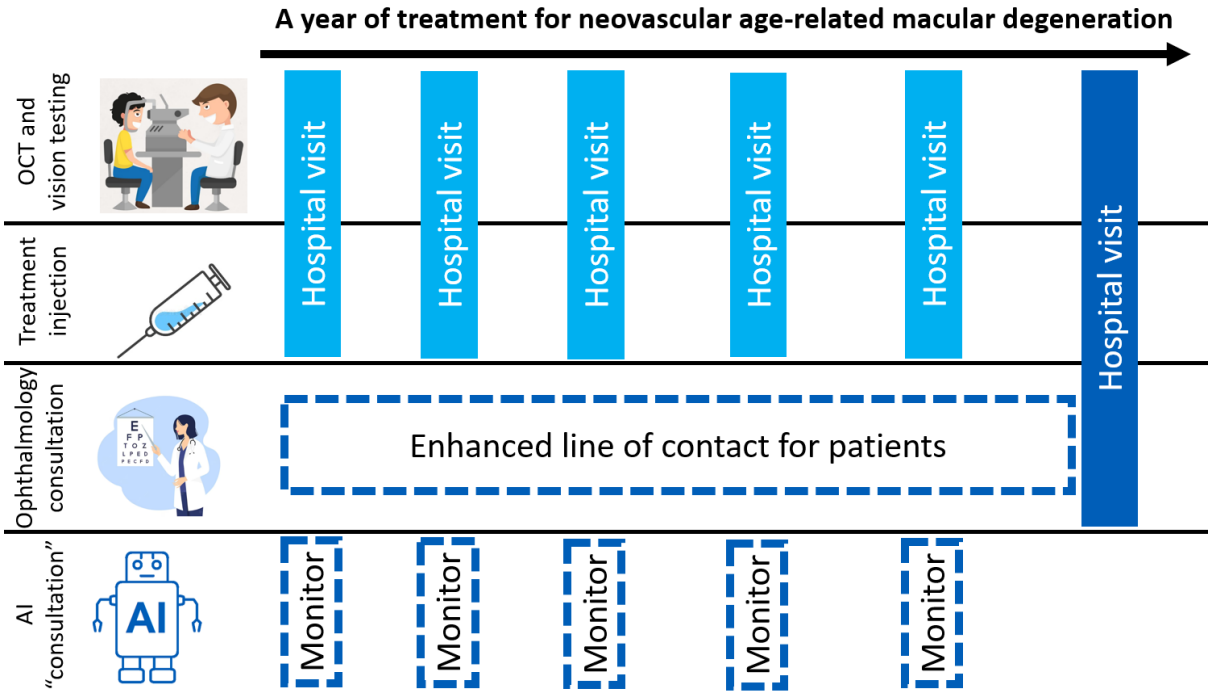


Figure 37. Schematic of proposed AI-enabled intervention for nAMD treatment monitoring.

6.7 Discussion

The multiple elements of a complex healthcare intervention, such as AI-enabled nAMD treatment monitoring, provide many mechanisms which could influence its implementation. Analysing determinants of implementation from chapter 3 with the FITT framework has offered insights into how an AI-enabled intervention for nAMD care might exploit these mechanisms to deliver value to key stakeholders.[21] The theory-informed analysis has also produced a transparent evidence base for each element, which can be iterated upon as the determinants underlying the rationale for intervention design vary over time and between implementation settings. This transparency lends resilience and generalisability to the value of the proposed intervention as the feasibility of design re-iteration by others with expertise in distinct implementation settings is improved. Given the strong influence context appears to have on implementation outcomes, and the lack of precedent for AI-enabled nAMD care, this potential for responsive re-iteration is likely to safeguard the proposed modifications to impact real-world care.

6.7.1 Comparison with prior work

Much of the published research to support clinical AI intervention design and evaluation falls within the overlapping fields of usability, ergonomics or human computer interaction.[3] These fields are also identified as producing a distinct group of TMFs according to one taxonomy.[28] Primary studies here often focus on potential users' interaction and experience with a clinical AI prototype.[3] It is also noteworthy that much of this work is done by vendors developing AIaMD and is not in the public domain, due to the resource demands of publication and the commercial sensitivity of its content. Prototype evaluation provides a close simulation of what the product might feel like and so qualitative data collected from study participants has a high degree of authenticity.[29] There are some disadvantages however, as these simulated usability studies often have a narrow focus on the AIaMD, rather than the intervention as a whole.[21] When they are not informed by prior explorative research, they can be indicative of a solution-first approach to innovation, rather than a problem-first approach.[5] This risks developers failing to anticipate certain demands of the AIaMD's target users or implementation context and limiting the effectiveness of the resources they commit to prototype development. Even if developers are willing and able to enact major changes to better align a prototype with these demands, it is likely to represent an inefficient development process. This solution-first approach is exemplified in a usability study of a virtual nasal surgery decision support tool which presents a relatively mature prototype to potential users and asks what task it could be used for.[30] Contrastingly, a separate investigation of an asthma decision support tool explored clinicians experience and need across the healthcare pathway using a TMF to help inform subsequent prototype development and evaluation (Figure 38).[31]

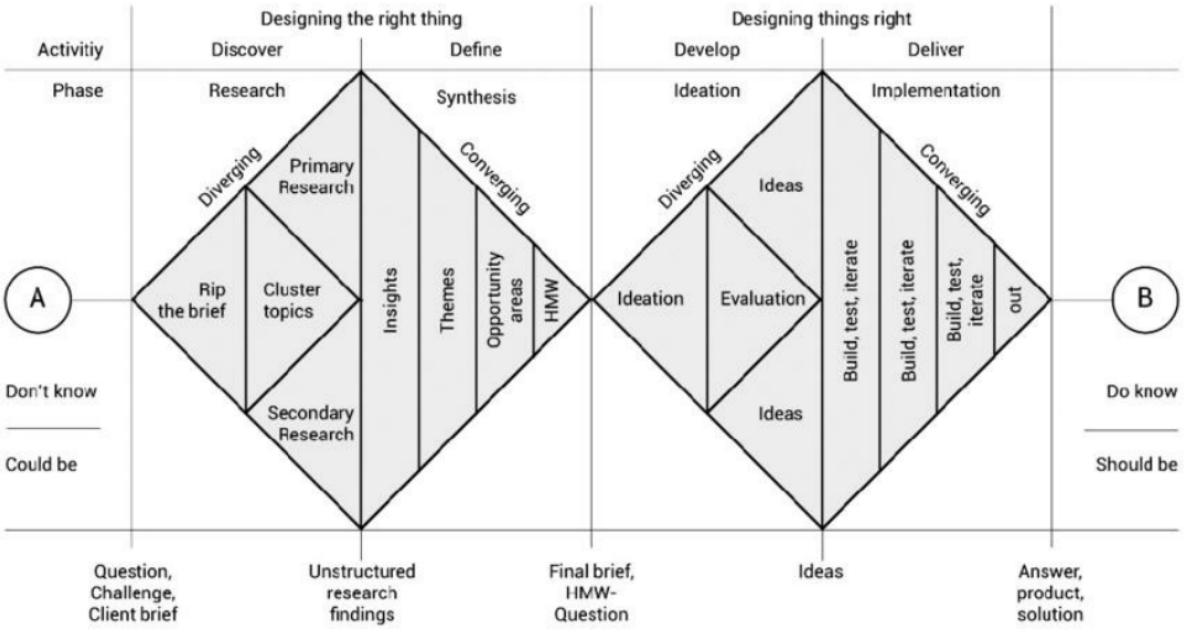


Figure 38. The universal double diamond framework, a two-stage design-framework focusing first on what kind of intervention is needed and then what form that intervention should take. HMW =How-Might-We questions[32]

Though outside of scope for the present study, greater authenticity in evaluating AI-enabled interventions is possible following their integration into live clinical workflows. This work remains uncommon, though is becoming more prevalent. With the present research, findings are highly practical as they are based on real-world experiences and events with clinical AI. One example of the value of such practical findings comes from the employment of nurse specialists to communicate outputs from an AI tool to end-user emergency department physicians.[33] Here the addition of a human interface for AI resulted in a relatively positive experience and high acceptance from end-users. These studies of live AI-enabled interventions can also highlight unanticipated risks which require mitigation. An example here was the observation of nurse specialists in a Thai community diabetic screening service making ad-hoc decisions to re-image patients if they disagreed with a decision for onward referral in an AI-enabled intervention.[34] This helped to surface training needs for end-users and other considerations about the suitability of implementation contexts. In the present study, the data on which the intervention design are based come from participants who have a hypothetical understanding of it. This was necessitated by the absence of a prototype at the time the study was developed, but also has the advantage of anticipating a proportion of these insights whilst there is still significant flexibility in prototype and intervention design.

A major strength of this work is the multi-stakeholder insights it draws on and the process of exploring compromises or alternate scenarios where conflict arises between those. This need to iteratively refine AI-enabled interventions through multi-stakeholder input was also well demonstrated in a qualitative study of stakeholders in a live integration of hospital bed occupancy predictive AIaMD in the US.[2] The data from this study were abstracted into a process model for iterative, collaborative development and implementation of clinical AI tools (Figure 39).

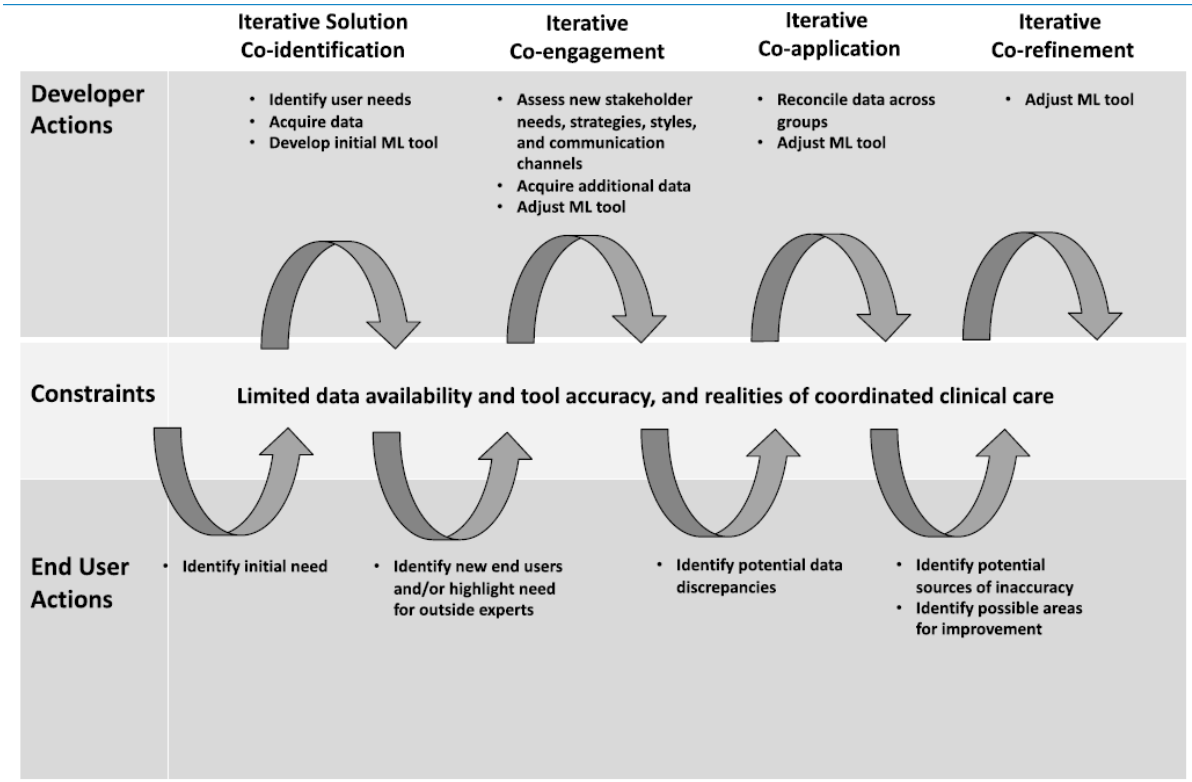


Figure 39. Iterative, collaborative development and implementation of machine learning-based clinical decision support tools.[2]

Similar sentiments were presented and abstracted in a primary qualitative study supporting the design of a AIaMD to consolidate unstructured data in patient records for nursing staff (Figure 40).[35] Both process models suggest that the work presented in chapter 4 and the present one will prove most successful as the initial stage of an on-going iterative co-development process.

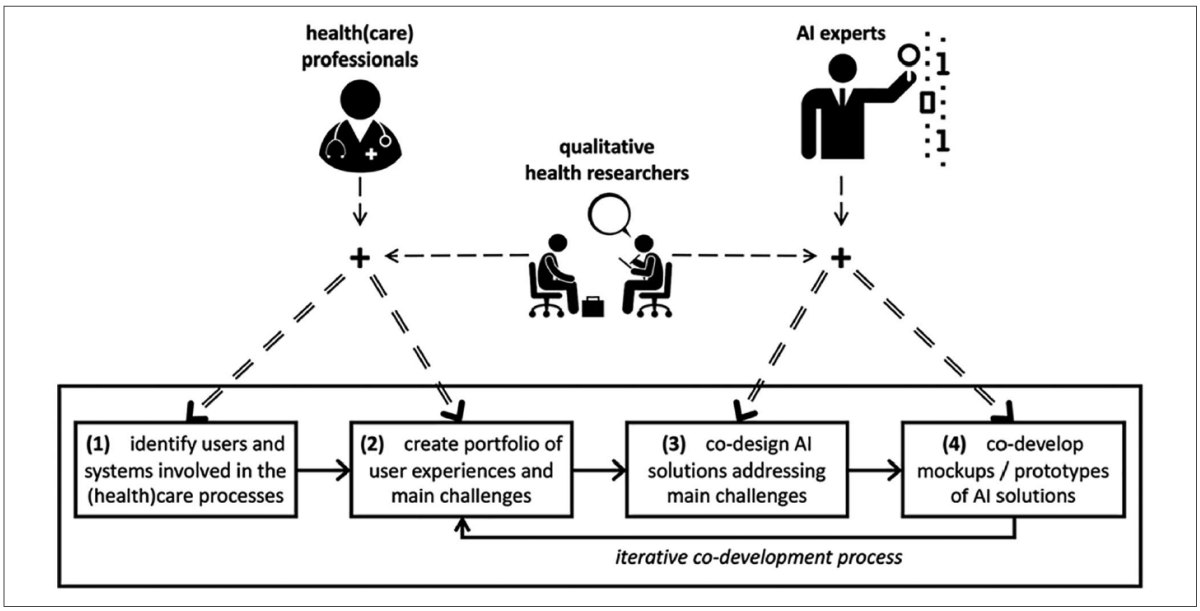


Figure 40. User-driven co-development of artificial intelligence model.[35]

6.7.2 Limitations

As discussed in the previous section, stakeholders' perspectives contributed to intervention design despite their lack of direct experience of that intervention. As a result, the authenticity of their insights are limited and they may experience the intervention differently once they are interacting with it in real-world care.[29] This is an inevitable limitation of this early stage hybrid research where the need to establish effectiveness must be prioritised over interventional implementation research.[13] Nevertheless, the observational form of implementation research here remains important to maximise the efficiency of further iterations on the intervention through simulation and interventional research (Figure 39 and Figure 40).[36]

Some of the decisions made through the FITT framework analysis were informed loosely by costs associated with the salaries of various staff members (Table 34).[21, 26, 27] To explore this more rigorously a formal health economic evaluation with sensitivity analysis should be performed.[12] This is outside the scope of the present work, but will be valuable in shaping the intervention and expressing the value proposition to adoption decision makers. There is another economic assumption which threatens the legitimacy of the proposed intervention. This relates to the 'time to care' value proposition, which is often assigned to AI by its advocates, suggesting that efficiency savings from AI will mean that clinicians have time to deliver higher quality encounters with patients when they do occur. This is reflected in the intervention proposed by the assumption that with fewer demands for face-to-face consultations in AI-enabled macula services clinicians will invest more time in each individual consultation. This helps to address concerns from clinicians described in chapter 4 that their consultations in macular services also allow them to make new diagnoses (e.g. cataract or glaucoma) or make more holistic interventions (e.g. arranging a low vision aid assessment or discussing visual driving standards). In reality, the capacity pressure that managers and clinicians are under may mean that any clinician time that is freed up is immediately assigned to the provision of more service, rather than enhancing the quality of service that is already provided. This can be mitigated against through implementation strategies, but the real-world delivery of 'time to care' should be monitored in any implementation efforts.

Another limitation of the theory-based design of the intervention is that it does not provide the level of specification required by some stakeholder groups in some regards. This limits the practical application of its current state. For example, regulators would require technical specifications of the type of OCT imaging that would be used as inputs into the AI device.[1] Prior to any implementation effort it will be necessary to establish these details, but the advantage of the present approach is the broad accessibility of the output it has generated from complex qualitative data. This provides a useful starting point to engage stakeholders in subsequent refinement of the intervention for specific settings.

Given the national, or international, scope of regulators of SaMD the local focus of data collection to support intervention design is another limitation. This is of particular concern given the low ethnic diversity in the locality studied (Northeast of England) and the well-placed concerns over potentially biased impacts of AI interventions across demographic groups.[37, 38] The local focus of this tool helped to provide common experiences of nAMD care. In so doing, it provided an opportunity for participants to draw on their lived-experiences when reflecting on the potential implementation and use of the AI-enabled

intervention. In further work however, the findings of the study must be critically evaluated with insights from a more diverse sample of patients, carers and clinicians.

6.7.3 Future directions

To support regulatory approval of an intervention such as the one outlined in the present chapter, further specification is required within the framework used by regulators and other decision makers. Chapter 7 will address this by applying the formal structure of an IUS, which is central to regulatory submissions for SaMD in most global jurisdictions and helps to provide practical clarification of the form of an intervention.[1]

To address the systems perspective recommended in the Medical Research Council's framework for evaluating complex interventions, it is also necessary to consider the proposed healthcare pathway in which the proposed AI-enabled intervention will sit.[11] This will be addressed through a MAA in chapter 7. This involves mapping a hypothetical AI-enabled healthcare pathway and evaluating the risks that may present at each stage.[39] A further error analysis will be performed as part of this too, to explore what mitigations may be available across the full care pathway, rather than just from the AI technology, AIaMD or AI-enabled intervention. This process will be used to make recommendations on opportunities to improve the safety and effectiveness of the proposed AI-enabled care pathway further.

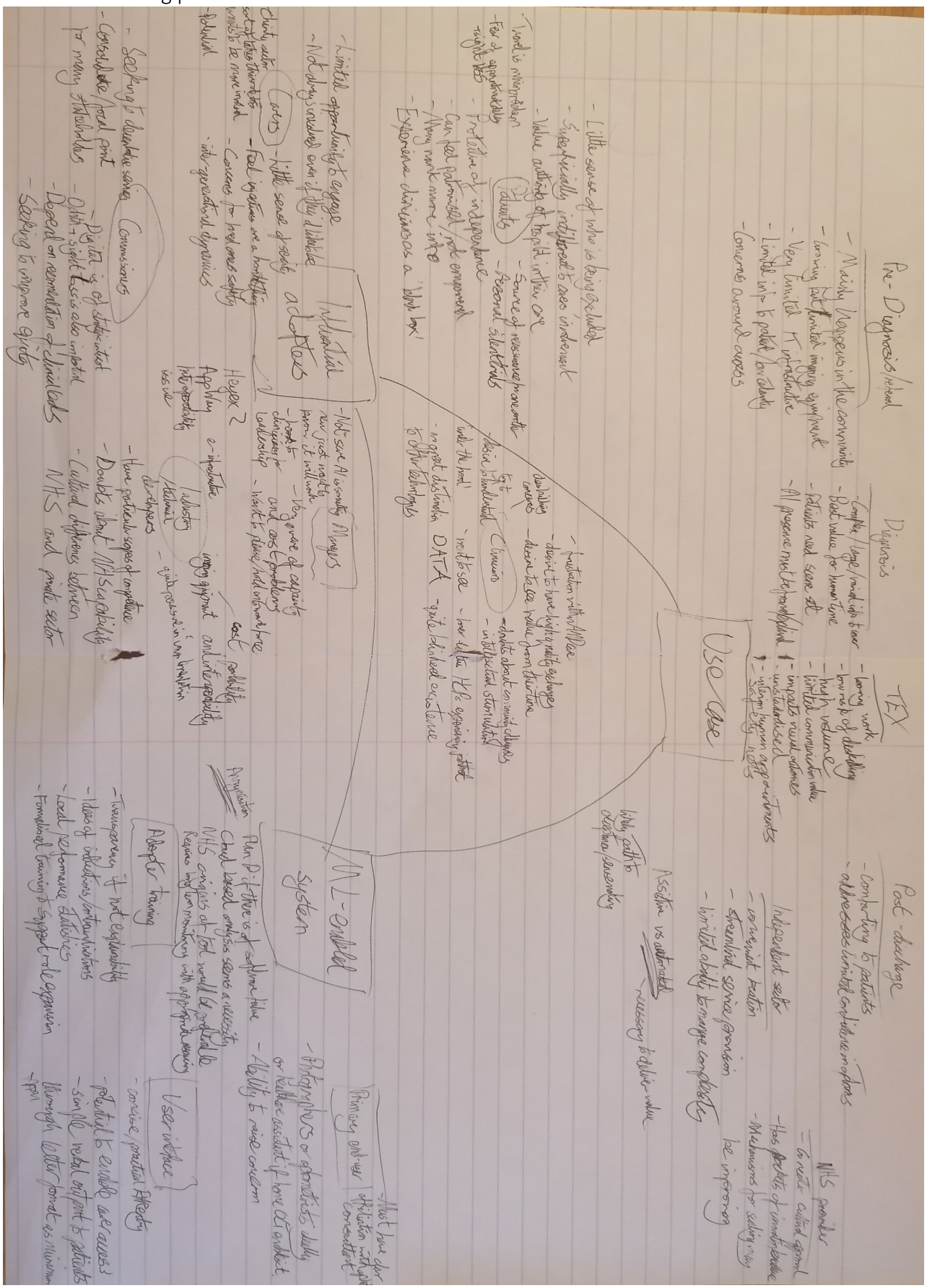
Addressing other limitations described in the prior section is outside of the scope of this thesis. However, the associated evidence will be needed by regulators and health technology assessors in their stop-go decisions about implementation in the UK and/or NHS. This includes formal health economic evaluation, qualitative research to explore perspectives from stakeholder at other sites across the UK and AIaMD prototype development and evaluation. Findings from these work streams will also help to evaluate and refine the intervention further and to prepare for implementation.

6.8 Conclusions

Drawing on the FITT Framework, this chapter has proposed a testable intervention with a transparent evidence-base. This takes the form of an alteration from the current nAMD treatment pathway to reduce the number and increase the quality of F2F consultations between clinicians and patients. This can be expected to improve patient and clinician experience whilst also increasing service efficiency. This acts as a starting point for review of a full AI-enabled healthcare pathway and its further refinement to improve the safety and effectiveness than can be expected from AI-enabled nAMD treatment monitoring.

6.9 Appendix

6.9.1 FITT drafting process



6.9.2 Potential patient leaflet

New technology in your macular degeneration clinic

We have been testing new technology against specialist consultants here and at Moorfields Eye Hospital to try and improve your experience at clinic. This new technology looks at the photos we take to check on your wet macular degeneration and carefully plan when your next treatment should be.

How do I know it is safe?

This technology has been developed by teams in the NHS, UK universities and companies over the last 7 years. Like the other technologies we use in your care it has been approved for use by the UK regulator, the Medicines and Healthcare products Regulatory Agency (MHRA). With the rest of the team, our consultants have also been making their own checks before using the technology in clinic. We also want to show you the treatment plans it makes at your appointments here at the clinic, so you can see it at work for yourself.

How can this technology benefit me?

At Newcastle Eye Centre we have shown that our patients can already expect better vision with wet AMD than the average for the UK. In checking this technology on our own medical records, we found that we could make this even better. The technology could make us better at catching wet macular degeneration getting worse and finding more opportunities to give injections less often without losing vision.

Most of the time spent waiting in clinic is for your team to look through your photos, so by bringing this technology in we expect you to spend less time waiting. This technology can also make use of photos taken in some community health centres too, so we are hoping to make it easier for you to get to places that we can give you care.

What if I think there is a problem with the technology?

Your safety and the safety of your vision is extremely important to us. We believe this technology will benefit the people with wet macular degeneration we care for, but we also understand that this is new and that you often know first when you need to be seen. To help we have also introduced a new phoneline for the wet macular degeneration clinic so that you can raise any concerns you have. A member of the team will call you back to discuss the concern, check your records and arrange an extra appointment if it would improve your treatment. This is in addition to our other services.

6.9.3 nAMD documentation quality improvement project

The Newcastle upon Tyne Hospitals NHS Foundation Trust

Neovascular age-related macular degeneration clinic letters: a quality improvement project

Melissa Gough, Jeffry Hogg, Vina Manjunath
Royal Victoria Infirmary, Eye Department, Newcastle

Introduction

Age-related macular degeneration (AMD) affects 196 million people worldwide (2020), making it the most prevalent retinal disease in the Western world[1]. Neovascular age-related degeneration (nAMD) comprises 10% of cases, with non-neovascular AMD comprising the other 90% of cases[2]. One arm of current treatment strategies comprises protocol-based anti-vascular endothelial growth factor (VEGF) injections. It is essential for optimum patient care and safety that treatment plans and disease course are communicated clearly in documentation and clinic letters so that they can be managed to the highest standards.

Methods

A randomised selection of nAMD patients (n=100) seen in macular clinic at the Eye Department, RVI, were sampled from Medisoft during September 2022. Data fields are summarised below (Fig 1). Data were analysed using descriptive methods and independent t-tests were used to compare means. An intervention was undertaken in the form of departmental teaching. Data collection was repeated in January 2023.

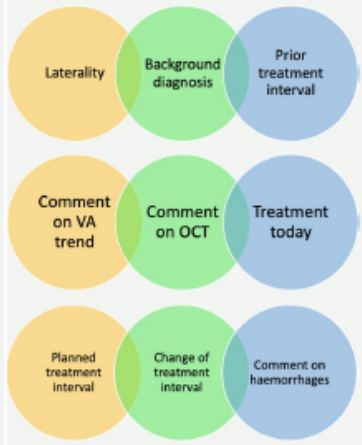


Figure 1: Data fields collected

Primary aim

- Check current data completeness in nAMD clinic letters.

Secondary aim

- Produce a gold standard letter template according to departmental consensus.

Results

Patient demographics were equivalent across both cycles (age cycle 1 mean 80, 55-97; cycle 2 mean 80, 39-97 *p=0.960*, months on treatment cycle 1 mean 47m, 0-133m); cycle 2 mean 43m, 0-136m *p=0.508*). Improvements in data completeness across all fields was seen after intervention was delivered. Discussion with departmental staff (consultant ophthalmologists and specialist ophthalmology nurses) highlighted that some data fields (diagnosis, number of treatments) are automatically recorded in Medisoft and pulled through into any letters generated. These data, therefore, are not necessary to record in every clinic note. The consultant survey revealed variable opinion regarding essential versus non-essential factors.

Specific Improvements noted between cycle 1 (Fig 2) and cycle 2 (Fig 4) were demonstrated in recording prior treatment (cycle 1 24% versus cycle 2 53%; 29% absolute increase) and change of interval (cycle 1 31% versus cycle 2 62%; 31% absolute increase).

Interestingly, consultant opinion (Fig 3) on the necessity to record comments on haemorrhages was variable and tended toward non-essential (essential 25% versus non-essential 75%). This could explain why the recording of this information remained low throughout both cycles.

Conclusion

We have shown that the use of a teaching intervention and distribution of a survey concerning letter information helps to improve data completeness in nAMD patient clinical notes and letters. This has positive implications for safety-optimised patient treatment.

Recommendations

- Organise forum for discussion concerning consensus for essential data
- Further data collection cycle to monitor improvements
- Plan any further necessary interventions arising from results of next cycle

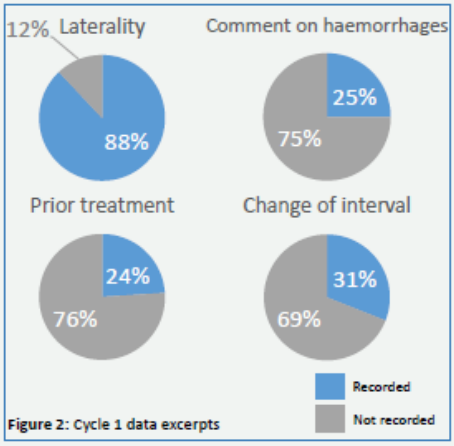


Figure 2: Cycle 1 data excerpts

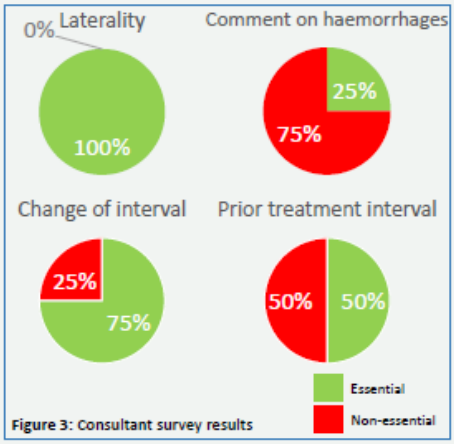


Figure 3: Consultant survey results

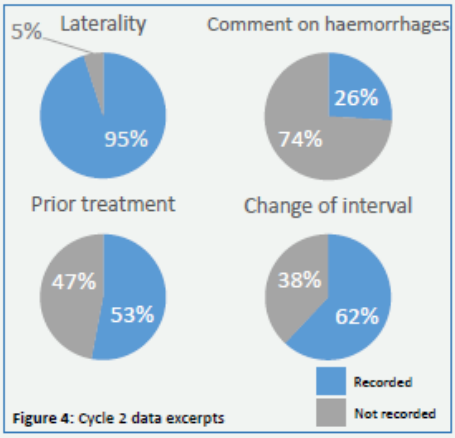


Figure 4: Cycle 2 data excerpts

References

1. Wong WL, Su X, Li X, Cheung CM, Klein R, Cheng CY, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Health. 2014;2(2):e106-16.
2. Ambati J, Fowler Benjamin J. Mechanisms of Age-Related Macular Degeneration. Neuron. 2012;75(1):26-39.

6.9.4 Photos from public engagement event

6.10 References

1. The Medicines and Healthcare products Regulatory Agency, Crafting an intended purpose in the context of Software as a Medical Device (SaMD). 2023.

2. Singer, S.J., et al., Enhancing the value to users of machine learning-based clinical decision support tools: A framework for iterative, collaborative development and implementation. Health Care Manage Rev, 2022. 47(2): p. E21-e31.

3. Hogg, H.D.J., et al., Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence. J Med Internet Res, 2023. 25: p. e39742.

4. Greenhalgh, T., et al., Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res, 2017. 19(11): p. e367.

5. Kim, J.Y., et al., Organizational Governance of Emerging Technologies: AI Adoption in Healthcare, in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023, Association for Computing Machinery: Chicago, IL, USA. p. 1396–1417.

6. Abràmoff, M.D., et al., Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine, 2018. 1(1): p. 39.

7. Ong, A., et al., Artificial Intelligence as a Medical Device for Ophthalmic Imaging in Europe, Australia, and the United States: Protocol for a Systematic Scoping Review of Regulated Devices. Open Science Framework, 2023.

8. Peter, H., et al., Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. British Journal of Ophthalmology, 2021. 105(5): p. 723.

9. Ipp, E., et al., Pivotal Evaluation of an Artificial Intelligence System for Autonomous Detection of Referrable and Vision-Threatening Diabetic Retinopathy. JAMA Network Open, 2021. 4(11): p. e2134254-e2134254.

10. Zhelev, Z., et al., Automated grading in the Diabetic Eye Screening Programme, U.N.S. Committee, Editor. 2021.

11. Skivington, K., et al., A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Bmj, 2021. 374: p. n2061.

12. Hull, L., et al., Designing high-quality implementation research: development, application, feasibility and preliminary evaluation of the implementation science research development (ImpRes) tool and guide. Implementation Science, 2019. 14(1): p. 80.

13. Brown, C.H., et al., An Overview of Research and Evaluation Designs for Dissemination and Implementation. Annu Rev Public Health, 2017. 38: p. 1-22.

14. Hogg, H.D.J., et al., Evaluating the translation of implementation science to clinical artificial intelligence: a bibliometric study of qualitative research. Front Health Serv, 2023. 3: p. 1161822.

15. Glasgow, R.E., et al, Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. Am J Public Health, 2003. 93(8): p. 1261-7.

16. Topol, E.J., High-performance medicine: the convergence of human and artificial intelligence. Nat Med, 2019. 25(1): p. 44-56.

17. Aristidou, et al, Bridging the chasm between AI and clinical implementation. Lancet, 2022. 399(10325): p. 620.

18. Camaradou, J.C.L. and H.D.J. Hogg, Commentary: Patient Perspectives on Artificial Intelligence; What have We Learned and How Should We Move Forward? Adv Ther, 2023. 40(6): p. 2563-2572.

19. Nilsen, P., Making sense of implementation theories, models and frameworks. Implementation Science, 2015. 10(1): p. 53.

20. Henry, K.E., et al., Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digit Med, 2022. 5(1): p. 97.

21. Ammenwerth, E., C. Iller, and C. Mahler, IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. BMC Medical Informatics and Decision Making, 2006. 6(1): p. 3.

22. Davis, F.D., Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly, 1989. 13(3): p. 319-340.

23. Sittig, D.F. and H. Singh, A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care, 2010. 19 Suppl 3(Suppl 3): p. i68-74.

24. Gance-Cleveland, B., et al., Using the Technology Acceptance Model to Develop StartSmart: mHealth for Screening, Brief Intervention, and Referral for Risk and Protective Factors in Pregnancy. Journal of Midwifery & Women's Health, 2019. 64(5): p. 630-640.

25. Benda, N.C., et al., “How did you get to this number?” Stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study. Journal of the American Medical Informatics Association, 2020. 27(5): p. 709-716.

26. National Health Service, Agenda for change - pay rates. 2023.

27. Mritish Medical Association, Pay. 2023; Available from: https://www.bma.org.uk/pay-and-contracts/pay.

28. Liberati, E.G., et al., What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci, 2017. 12(1): p. 113.

29. Hogg, H.D.J., et al., Unlocking the potential of qualitative research for the implementation of artificial intelligence-enabled healthcare. Journal of Medical Artificial Intelligence, 2023. 6.

30. Vanhille, D.L., et al., Virtual Surgery for the Nasal Airway: A Preliminary Report on Decision Support and Technology Acceptance. JAMA Facial Plast Surg, 2018. 20(1): p. 63-69.

31. Overgaard, S. Implementation, usability, and workflow integration of an ML-based CDS tool: A qualitative evaluation of user requirements. in American Medical Informatics Association 2022 Clinical Informatics Conference. 2022. Houston.

32. Jilka, M., Application Of The Double Diamond Framework To Prepare The Communication Strategy Of A Great Sports Event. Studia sportiva, 2019. 13(1).

33. Sandhu, S., et al., Integrating a Machine Learning System Into Clinical Workflows: Qualitative Study. J Med Internet Res, 2020. 22(11): p. e22421.

34. Beede, E., et al., A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020, Association for Computing Machinery: Honolulu, HI, USA. p. 1–12.

35. Schneider-Kamp, A., The Potential of AI in Care Optimization: Insights from the User-Driven Co-Development of a Care Integration System. Inquiry, 2021. 58: p. 469580211017992.

36. Curran, G.M., et al., Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care, 2012. 50(3): p. 217-26.

37. Ofice for National Statistics, Population estimates. 2021.

38. Suresh, H. and J. Guttag, A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, in Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. 2021, Association for Computing Machinery: --, NY, USA. p. Article 17.

39. Liu, X., et al., The medical algorithmic audit. The Lancet Digital Health, 2022. 4(5): p. e384-e397.

Chapter 7: Evaluating a proposed AI-enabled intervention for nAMD treatment monitoring

Problem: The risks and potential mitigations for a real-world AI-enabled nAMD treatment pathway are yet to be systematically explored. This increases the chances of unanticipated harm from the AI-enabled intervention following its implementation, but also the chances of research waste if regulators or other decision makers identify unanticipated risks and halt the implementation process.

Objectives: Apply the "Medical Algorithmic Audit" (MAA) to the AI-enabled pathway that would result from the intervention described in chapter 5, supported by the qualitative and quantitative data and findings of chapters 3 and 4. Use the findings of the MAA to refine the likely safety profile of the proposed AI-enabled intervention.

Methods: As far as possible, the six stages of the MAA will be completed for the proposed AI-enabled intervention in its current state of maturity (i.e. disjointed deep learning-enabled tissue segmentation with rule-based decision tree overlay). This will allow the data from chapters 5 and 6 to directly inform the Scoping, Mapping, Artifact collection, Testing, Reflection and Post audit stages of the audit.

Findings: The AI-enabled intervention is described in an Intended Use Statement in the form recommended by the MHRA. Further adjustments to the rule set applied to OCTane outputs can be applied to improve its NPV from 95.3% to 96.5% and PPV from 42.3% to 55.5%. Of the remaining 53 false positives, 24 (45.3%) are associated with major segmentation errors and of those 24, 22 (91.7%) are associated with suboptimal OCT quality. These observations could inform the design of user training, to minimise the risk of errors from low quality imaging and to increase the proportion of errors identified and averted by clinicians. Patients with certain ocular characteristics (myopic fundus, large pigment epithelial detachment or ocular media opacification) may experience lower diagnostic accuracy of AI-enabled decisions and further investigation of these sub-groups is warranted.

Conclusions: The MAA has facilitated the identification of risks and mitigations across a hypothetical AI-enabled nAMD treatment pathway. This process has helped to further optimise an AI-enabled intervention for prospective evaluation in nAMD treatment monitoring

7.1 Background

In chapter 4 we found that both the performance of AI technology and the manner in which it is applied to nAMD treatment monitoring will be two key determinants of implementation outcomes. We also learnt that both factors were relatively modifiable compared to other influential factors. Chapter 5 suggested that the AI technology under study would satisfy most stakeholder's expectations of performance when combined with certain knowledge-based rules within an AIaMD. Chapter 6 explored other aspects of that AIaMD and how the intervention to carry it could best align with determinants of implementation. So far then, we have considered the AI technology, the device which it may sit within and the intervention which might carry it into patient care.

This chapter will take another step out into the target context by considering its place within the wider hypothetical AI-enabled healthcare pathway. The intervention for AI-enabled nAMD treatment monitoring involves many different interdependent factors (e.g. adopters, technology, organisations, policy etc.) which influence one another in a non-linear or partly stochastic fashion. Such influences characterise complexity, as opposed to complicatedness (many factors influencing each other in a predictable way) or simplicity (few factors influencing each other in a predictable way).[1] The Medical Research Council has recently updated its framework for the evaluation of complex healthcare interventions to move beyond efficacy (performance in ideal settings) and effectiveness (performance in real-world settings). It now emphasises the value of evaluating the mechanism through which interventions achieve their real-world outcomes and the wider impact they have across the system they act within. Considering the interaction of the proposed AI-enabled intervention with the target healthcare pathway will improve the identification and accommodation of the diverse sociotechnical factors which will shape the outcome of implementation.[2, 3]

For individuals, organisations and systems across healthcare, the risk of causing harm to patients is the central constraint for clinical AI implementation. In some regards, this constraint is a self-perpetuating one as without real-world clinical AI implementation we can only estimate these risks and harms. Even in cases where clinical AI has made it to real-world use, the detection and reporting of risk and realisation of harms is minimal.[4] Even if purely technical failures of AIaMD are considered, a diagnostic device with a sensitivity of 80% (e.g. ProstatID™) should be expected to miss at least 20% of cases in practice.[5] The discord between the expected and reported degrees of risk and harm is not surprising, but it does highlight the limitation of current reporting practices in AIaMD post market surveillance.[6] A recent synthesis from post market surveillance reports submitted to the FDA suggested that most harms seem to be derived from the interactions that users have with AIaMD, rather than failures of either the users or products in isolation.[4] This emphasises the importance of approaches to risk assessment that consider AI-enabled healthcare as holistically as possible.

7.2 Problem

There is a widespread cultural scepticism about clinical AI across healthcare and from many stakeholders in nAMD treatment monitoring.[1, 7] To satisfy adopters, clinical AI must match and preferably exceed the effectiveness and safety of current service provision. Stakeholders' definitions of 'effectiveness' and 'safety' are multi-dimensional and can only be fully measured with varied quantitative and qualitative data. There are also no real-world instances of AI-enabled nAMD monitoring in the NHS, so approaches to evaluating

effectiveness and safety in the present work need to retain value even in hypothetical applications.

Many systematic or theory-informed approaches for hypothetical safety evaluation exist, but they are often at a high level of abstraction.[8-10] This makes it difficult to distil actionable findings for a specific application such as nAMD treatment monitoring. There are also instances of approaches which are focused enough on an application to produce actionable findings, but hold little intuitive relevance for clinical AI, limiting their value.[11] This lack of domain focus is problematic given the relative distinctions of AI against other health technologies of relatively opaque mechanisms of action and the potential to make complex decisions with high levels of independence.

To evaluate an AI-enabled healthcare pathway in a way that is meaningful for stakeholders and supports actionable recommendations a multi-method, clinic AI-specific approach is required.

7.3 Rationale

The MAA is a systematic approach to monitoring and safety evaluation which has been specifically designed for AIaMD following their integration into clinical practice.[12] As such it includes steps which specifically address some of the distinctions of AIaMD; subgroup and adversarial testing to compensate for the opacity of operating mechanisms, rigorous IUS for clarity on the contributions the AIaMD will make to care. It requires consideration of the full healthcare pathway, including elements which do not directly involve the AIaMD, to identify and mitigate risks arising from both the AIaMD and the wider context in which it will sit.[4] It also demands varied quantitative and qualitative inputs to holistically describe the performance of an AI-enabled intervention.

Although MAA was designed with longitudinal monitoring of AIaMD already incorporated into healthcare pathways, there is also precedent for it being used in formative evaluations of AIaMD prior to clinical integration.[12] Here it produced actionable recommendations to guide a potential AI-enabled healthcare pathway for diagnosing neck of femur fractures on plain radiographs. Consequently, MAA addresses the problems which this chapter aims to address.

7.4 Aim

To conduct a MAA of the proposed AI-enabled nAMD treatment monitoring pathway to systematically identify clinical risks and potential means for their mitigation.

7.5 Methods

The MAA was conducted in line with the steps outlined in the index publication.[12] A prior published example and an internal example used in governance processes at an NHS foundation trust were used as exemplars.[13] Not all steps of the MAA are possible, or require adaption, for an AIaMD which is yet to be integrated into practice such as the one proposed. These adaptions or omissions are explicitly called out in the relevant part of the results section below.

7.6 Results

7.6.1 Scoping

Because of the UK implementation context under consideration, the scoping stage draws on chapters 3 and 4 and guidance from the MHRA on the form and content of IUS.[14]

7.6.1.1 Structure and function of the device

This SaMD aims to support treatment monitoring for patients with an established diagnosis of nAMD. It achieves this by assessing whether signs of nAMD disease activity are present or not to inform the interval at which a suitably qualified clinician (e.g. ophthalmologist, optometrist or nurse practitioner) plans further anti-VEGF treatment.

The SaMD has 3 components; an application programming interface (API), a cloud-based deep learning algorithm (OCTane) and a rule-based algorithm.

Inputs to the SaMD are ipsilateral pairs of temporally sequential, topographically-matched, fovea-centred 6mm x 6mm OCT scans composed of 25 or 50 horizontally orientated, equally spaced B-scans. Prior to SaMD use, patients' OCT scans are stored by the data controller organisation in their PACS as DICOM files. To apply the SaMD, a pair of DICOM files from a patient are exported by a suitably qualified operator (e.g. medical photographer) to a cloud-based platform. Prior to this export the DICOM files are pseudonymised and encrypted within the data controller's digital environment. From the cloud-based platform, an application programming interface (API) within the SaMD strips all fields of the DICOM files besides the imaging itself prior to analysis by OCTane. OCTane has a three-dimensional U-Net architecture and segments full OCT volume scans into 15 different tissue groups, including IRF and SRF, assigning a total volume to each in μm^3. A rule-based algorithm then applies simple logic to differences in IRF and SRF volumes between the paired OCT scans, determined by OCTane. If either SRF or IRF volumes increase by more than 1,000,000 μm^3 from the earlier scan to the later scan, then this confers an output of disease activity for that patient episode. If neither IRF or SRF volumes increase by 1,000,000 μm^3 or more, an output of disease stability is assigned for that patient episode.

Alongside this binary assessment of nAMD disease activity, the original raw OCT scans, graphical representations of the OCTane scan segmentations and the scalar volumes of IRF and SRF assigned to each are written to a DICOM file. This DICOM file is returned to the cloud platform by the API. These outputs are then exported back to the data controller's PACS where the encryption and pseudonymisation is reversed prior to storage in the relevant patient file, time and date stamped. The outputs can be accessed directly from the PACS or an interoperable EMR in the data controller's digital environment to support suitably qualified clinicians in deciding when they or a colleague should administer further anti-VEGF treatment to the patient. The raw OCT images, OCTane segmentations and numerical trends in IRF and SRF are presented alongside the binary assessment of disease activity to provide explainability.

7.6.1.2 Intended population

This SaMD is intended to be used for patients with an established diagnosis of nAMD who have given informed consent for a course of anti-VEGF treatment to the affected eye(s).

7.6.1.3 Intended user

There are a number of different users of the SaMD; the clinician conducting F2F consultations with patients, the clinician exporting OCTs for SaMD analysis, the clinician making nAMD treatment decisions and the patient.

The clinician conducting F2F consultations with patients is responsible for explaining the SaMD-enabled care pathway to patients with known nAMD. They should also address ad hoc communications from patients or colleagues who raise concerns about SaMD performance. On a regular basis (e.g. annually) they should conduct consultations with patients to ensure appropriate SaMD performance between consultations and check patients' understanding of diagnosis and consent for the on-going management plan. These consultations also allow the clinician to elicit patient concerns, screen for ocular comorbidities, social and functional consequences and answer any relevant questions that patients have.

The clinician exporting OCT images is responsible for checking that the patient has an active assignment to AI-enabled treatment monitoring for nAMD from a suitably qualified clinician. They are also responsible for ensuring that the correct pair of OCT images are uploaded and a technical quality check of the raw OCT images, as segmentation and decision support errors are more prevalent in low quality OCTs. After they export the OCT images they should check that the outputs are returned to the PACS as intended.

The clinician making nAMD treatment decisions also holds responsibility for checking that the SaMD outputs they access relate to the intended patient and imaging episodes. They alone are responsible for checking that the graphical displays of the segmentation appear accurate and that the numerical trends in IRF and SRF are also congruent with the raw and segmented OCTs. Assuming each of those checks is satisfactory they can then apply the binary assessment of disease activity output from the SaMD to local treatment protocols. This also requires knowledge of when the patient's last anti-VEGF injection was and potentially how long they have been treated for. They should also synchronously or asynchronously communicate their SaMD supported assessment of nAMD disease activity to the patient. This should be accompanied by the planned interval to their next treatment and a reminder of means by which patients can contact the clinical team with concerns regarding treatment. They should complete any prescribing or administrative tasks which enable the timely administrating of the next planned anti-VEGF treatment.

The patient has no responsibilities, but they are enabled to raise concerns about the treatment decisions or their symptoms directly to clinicians they interact with or through specified communication channels with the clinical team.

7.6.1.4 Intended use environment

The digital environment for SaMD analysis is a cloud platform. A PACS system, potentially accessed via an interoperable EMR, within the data controller's own computer environment is intended for the storage and access of SaMD outputs.

F2F consultations with patients are intended to be conducted within a clinical setting allowing for confidential discussion and ophthalmic examination. This could be in a primary or secondary care setting. The exporting of OCTs to the cloud environment for SaMD analysis and the accessing of outputs and clinical decision making could happen in the same

setting, or an office setting. This office could also be on site at the primary or secondary care premises or via secure remote access depending upon the data controller's digital infrastructure and information governance policies.

7.6.1.5 Intended impact

This intervention is intended to reduce the amount of clinician time required for nAMD treatment monitoring. It is intended to preserve or improve visual outcomes form nAMD patients undergoing treatment, without increasing the number of annual injections required by a patient. The intervention should also facilitate longer F2F appointments for nAMD diagnosis discussion and annual review consultations, with the expectation of improving patient and clinicians' experience.

7.6.2 Mapping

Given the pre-implementation stage of the intervention this mapping process is necessarily hypothetical. However, the stakeholders and data from chapter 3 help to inform this.

7.6.2.1 Personnel and resources necessary for audit

From NuTH, various different staff groups are required. A senior ophthalmologist is required to evaluate the OCTane segmentations and clinical decisions that were based upon those segmentations. An administrative staff member is required to curate data on patient appointment bookings, the rate of attendance and delay and capacity to provide different appointment types. A technical member of staff is required to evaluate the performance of the digital infrastructure on which the pathway depends. Following implementation, a sample of patient and professional end-users would also be surveyed or interviewed to understand their experience.

External staff from the SaMD vendor are required to provide insight into data flow and processing through the pathway. Assuming there is a different vendor for the cloud platform, PACS and imaging equipment, input will also be required from their teams. Independent benchmarking of the clinical cases to be examined in the 'testing' section can also be supported by an external reading centre or recognised specialist centre.

7.6.2.2 Risks and known vulnerabilities

7.6.2.2.1 Mapping of the AI system

The first dependency of the AI system is the capture of fovea-centred OCT imaging with equipment and imaging protocols that are interoperable with OCTane. This can also be affected by clinical factors, e.g. patient co-operation or optical media opacity, that affect the quality of the imaging obtained. Because the interpretation of the AI system's outputs is temporally dependent, the accurate time stamping of images in the PACS is also important.

Once the imaging has been stored in the PACS, the operator needs to select the appropriate AI tool and initiate its application (Figure 41). This is followed by image pseudonymisation and encryption prior to export to the cloud platform. Export depends on intact cloud infrastructure within the data controller's computer environment. AI analysis within the cloud environment takes place next and is dependent upon the interoperability of the imaging file format (DICOM).

The first stage of AI analysis is segmentation of the pair of OCT volume scans. The accuracy of this segmentation appears to be dependent on the quality of the imaging captured, but

may also vary across clinical groups, e.g. high myopes or individuals with extensive pigment epithelial detachment. Following segmentation, the differences between volumes of IRF and SRF between the paired scans is then derived by subtraction. If there has been an interval increase in either IRF or SRF volume over 1,000,000 μm^3 then an output of nAMD disease activity is assigned, otherwise an output of nAMD disease stability is assigned.

These outputs of visualisations of the segmentations, numerical assessments of IRF and SRF and the binary assessment of nAMD disease activity are then returned to the cloud environment and on to NuTH's PACS where the pseudonym key is reversed to return patient identifiers to the outputs. Clinicians can then access these outputs when making decisions about the relevant nAMD patient's treatment to integrate the AI outputs into the care pathway.

Figure 41. Diagram provided by Heidelberg Engineering™ to demonstrate data flows through Appway™, an example cloud-based platform to interface between imaging data archived in a Picture Archive and Communications System (PACS) and the AI tool.

7.6.2.2.2 Mapping of the health-care task

The task which the AI tool supports is determining the time at which patients with a known diagnosis of nAMD should receive their next anti-VEGF IVI. In the proposed AI-enabled pathway there are two appointment types; annual F2F review appointments and injection only appointments (see appendix). Due to the frequency of IVT treatment (Table 18), patients could expect to have 5 injection only appointments and 1 F2F review appointment over a year.

The annual review appointment would be very similar to the current pathway, refracted VA by an optometrist would be a routine part of the visit. There would also be more time afforded to the F2F consultation to ensure that patients have time to raise any questions or concerns and that clinicians have time to ensure understanding about diagnosis, management and prognosis. Another addition throughout treatment is an established channel for patients to raise concerns from home through administrative staff, to be forwarded to clinicians for asynchronous reply as needed.

In the injection only appointments patients will receive an OCT and an injection and then go home expecting to receive a letter commenting on the stability or progression of their

disease and a time for their next appointment. Separately from the patient's journey, the photographer will apply the AI tool to the OCT imaging and ensure the AI outputs are recorded within the PACS and EMR. In virtual clinics, clinicians will then review the EMR and PACS of nAMD patients who have attended to screen the AI segmentations and IRF/SRF quantifications for plausibility. They will then interpret the AI-enabled binary assessment (or their own independent assessment where AI errors are suspected) of disease activity into a recommendation for when the next IVT should be given. This will be forwarded to administrative staff who will book this as a F2F or injection only appointment. The nature of the appointment will depend on the time since their last F2F review, and a letter reflecting this will be sent to the patient.

7.6.2.2.3 Risk mapping

Risks are attributable to each element of the healthcare task and AI system above. To assign a relative priority to these risks to pursue mitigation a tripartite scoring system was applied (Table 36). Four of the risks with the highest priority score (>6) were pre-existent components of the nAMD treatment pathway and three related to the AI system (see appendix). The highest risk priority score to be attributed was ten, which was only attributed to the risk associated with poor quality imaging. This was assigned a severity score of 3, was demonstrably common from chapter 4 (assigned a 4 for occurrence) and requires thorough examination to detect on a single B-scan (assigned a 3 for detection).

Table 36. Risk priority scoring system

Score component	Numerical range	Practical definition
Severity (severity of failure effects)	1 2 3 4	No harm or anxiety/inconvenience >1 clinically unnecessary injections/year Vision loss with sequential occurrence Vision loss with one-off recurrence
Occurrence (likelihood of occurrence)	1 2 3 4	< 0.1% patient years 0.1% - 1% patient years 1% – 10% patient years > 10% of patient years
Detection (effectiveness of error detection mechanisms)	1 2 3 4	1 or more dependable detection mechanism 50% of errors expected to be detected 10 – 50% of errors expected to be detected <10% of errors expected to be detected

7.6.3 Artefact collection

Proxies for many of the necessary artefacts for an algorithmic audit are available (Table 37). For audits of live AI-enabled care pathways, documents which more directly represent the desired artefacts will be necessary.

Table 37. Artefact checklist for algorithmic audit. IUS = Intended Use Statement, SaMD = Software as a Medical Device, FMEA = Failure Modes and Effects Analysis, PACS = Picture Archiving and Communication System

Artefact	Present availability	Desired availability
IUS	Proposed IUS based in chapters 4 and 5	IUS against which SaMD regulatory approval was given
Intended impact statement	Value proposition derived from chapter 3	Local impact statement collectively established and revised by clinical, operational, technical and patient community leads
FMEA clinical pathway mapping	Swimlane diagram of proposed workflow derived from chapter 4	Swimlane diagram of workflow observed in real-world AI-enabled pathway
FMEA risk priority number document	Assigned by auditors on mix of empirical and anecdotal evidence	Local clinical, operational, technical and community leads to assign scoring framework and observed empirical basis for each score
Datasets	Training and validation datasets for OCTane are not available Rule-based decision tree is clinically based and validation dataset is available	Training and validation datasets for OCTane Multi-centre validation dataset for rule-based decision tree
Data description	Descriptives for training and validation datasets for OCTane are published [15] Descriptive for the rule-based decision tree validation is available	Descriptives for training and validation datasets for OCTane Descriptives of a multi-centre validation dataset for rule-based decision tree
Data and associated SaMD outputs for direct assessment	Retrospective, local, randomly sampled dataset available with all outputs available	Prospective, contemporary local dataset with all outputs available and oversampling of ‘edge cases’

Data flow diagram	Low detail data flow figure from potential PACS vendor	Detailed and granular dataflow diagram reflecting the reality of local digital infrastructure
AI model code	Not available for OCTane Rules for decision tree and alternatives are fully available	Availability for SaMD is unlikely as it will be commercially sensitive assuming that the vendor is distinct from the provider Rules for decision tree and alternatives
Model summary	OCTane architecture available in a research publication [15] Rules for decision tree and alternatives are fully available	Structured documentation of key model components [16]
Previous evaluation materials, performance testing and user experience	Chapter 3 and 5 represent qualitative and quantitative evaluations of a potential intervention Evaluations conducted during model development and internal validation have also been published	Reports from prior audits conducted at the local institution and other healthcare providers (redacted where necessary) to include survey, focus group or interview feedback from users

7.6.4 Testing

7.6.4.1 Exploratory error analysis

Following error analysis in chapter 5, four new rule-sets were iteratively explored, R8 – R12, using absolute thresholds for IRF and SRF as opposed to proportional threshold. These thresholds were increased stepwise in 1,000,000 μm^3 increments until an increase in the number of FNs was observed. The number of FNs (N=262) increased from 5 to 7 when the IRF threshold was lifted to 2,000,000μm^3 and from 5 to 6 when the SRF threshold was lifted to 3,000,000μm^3. With each elevation the number of FPs decreased and so rule set 10 (IRF threshold of 1,000,000μm^3 and SRF threshold of 2,000,000μm^3) appeared to offer the best performance (Table 38).

Table 38. Comparison of diagnostic accuracy statics between consultant-led care (CLC) and rule sets (R) derived from the pilot dataset, initial exploration of the full dataset and subsequent exploration informed by error analysis. NPV = Negative Predictive Value, PPV = Positive Predictive Value, CI = Confidence Interval

	NPV (95% CI)	PPV (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
CLC	80.8% (71.3-90.4)	42.2% (30.3-54.2)	53.5% (41.4-65.6)	72.8% (62.0-83.6)

R2	85.5% (77.0-94.0)	44.5% (32.5-56.6)	69.0% (57.8-80.2)	68.1% (56.8-79.4)
R6	95.3% (90.1-100.4)	42.3% (30.3-54.3)	93.0% (86.8-99.2)	52.9% (40.8-65.0)
R10	96.5% (92.1-101.0)	55.5% (43.4-67.5)	93.0% (86.8-99.2)	72.3% (61.4-83.1)

In applying R10 to the dataset of 262 cases there were 5 FNs which were the same cases that R6 assessed as FN discussed in detail in chapter 5. R10 also made 53 FPs in assessing the 262 cases, which compares favourably to the 90 FPs made by R6. Only 50 FPs were common cases between the two rule sets, R10 made 3 FPs for which R6 produced true negatives and R6 made 40 FPs for which R10 produced true negatives. Examining each of the 53 FPs made by R10 they appeared to fall into 3 causative groups; major segmentation errors, minor segmentation errors and clinically ambiguous cases (Table 39). Major segmentation errors were immediately apparent on glancing through the cartoon segmentations (Figure 32). On the majority of these cases (22/24, 91.7%) it could also have been apparent to the photographer that the quality of the image was suboptimal. Minor errors would not be apparent on such brief reviews of cartoon segmentations, but a proportion may be detectable by photographers reviewing the quality of images taken for AI analysis (8/9, 88.9%). CLC delivered true negatives (19/24, 79.2%) with a much higher frequency in the cases receiving R10 FPs due to the distinct mechanisms of failure between CLC and R10 (Table 39). This would suggest that any adaption to the intervention which signalled to clinicians that they should make independent decisions in cases of poor image quality or major segmentation error could improve the PPV.

Table 39. Categories and frequencies of error modes in false positives from rule-set 10. CLC = consultant-led-care

False positive error mode (n=53)	n (%)	n CLC true negative (row %)	low image quality (cropping, luminence or noise) n (row %)
Major segmentation error	24 (45.3%)	19 (79.2%)	22 (91.7%)
Minor segmentation error	9 (17.0%)	8 (88.9%)	8 (88.9%)
Clinically ambiguous	20 (37.7%)	8 (40.0%)	3 (15.0%)

It was notable that there were two examples of clearly myopic fundi among the 53 FPs. It is hard to assess if this is what would be expected from errors being made at random between patients with myopic and non-myopic fundi, but may warrant further investigation into performance within a myopic subgroup. Similarly, 7 of the 24 major segmentation error FP cases (29.2%) with major segmentation errors involved retinae with large domed pigment epithelial detachments, whilst just 2 of the 20 clinically ambiguous FP cases (10%) had this feature. Also relevant for further analysis, the OCTane tool was designed to analyse volume

scans made of 50 b-scans and consequently performs a pre-processing step on the 25 b-scan NuTH scans to double each b-scan. For the b-scans where major segmentation errors were made, there is a large amount of variation between the two attempts, whereas they are very similar in examples of good quality segmentation.

7.6.4.2 Subgroup testing

Subgroup analyses in all known demographic and clinical characteristic groups were performed to check for unequal distribution of FPs and FNs by R10 (see appendix). Alongside descriptive statistics, for categorical variables Chi-squared tests were applied to check for significantly different error distributions, whilst for continuous variables, independent two-sided t-tests were used. Whilst this represented significant multiple testing (32 statistical tests) there was no suggestion of unequal performance from R10 across the groups represented in the test dataset. Of particular note, prior discussion with clinicians and patients identified concerns of different error profiles among patients with macular haemorrhage. Of the 6 clinic visits where CLC included the documentation of a new or evolving macular haemorrhage, R10 correctly identified disease activity or stability in all cases.

7.6.4.3 Adversarial testing

Adversarial testing was outside of the scope of the present evaluation but would be recommended. Drawing on the exploratory analysis and clinical insights it would be useful to examine performance on a dataset enriched for patients with ocular media opacity (e.g. cataract or posterior capsule opacification), myopic fundi and neovascular disease (e.g. myopic macular degeneration) or patients with large domed pigment epithelial detachments as these features seem to promote major segmentation and disease activity assessment errors. Poor quality OCT imaging with low luminance or poor stability between B-scans would also be of interest, but harder to search for retrospectively.

7.6.3 Reflection

The retrospective evaluation of the initially proposed AI-enabled treatment monitoring device (with R6) demonstrated a NPV and PPV non-inferior to consultant-led care and perhaps superior when fully automated. This MAA has qualitatively highlighted that wider risks in an AI-enabled nAMD care pathway are largely comparable to those in the current pathway, but there are risk foci across the pathway which may be mitigated through developer and clinician actions.

7.6.3.1 Developer actions

In the final interface of the AIaMD containing this technology it would be valuable if a scrollable presentation of the cartoon segmentation was presented to clinicians. This could help them rapidly assess if there were major segmentation errors and whether or not the outputs were trustworthy. From the apparent lower precision of segmentation outputs on B-scans that cause errors, there may be a function by which the AIaMD could flag segmentations which clinicians ought to be suspicious of. This could be by performing multiple segmentations on each b-scan and seeing if the volumes of IRF and SRF returned demonstrate low or high precision. This may introduce disadvantages in the compute and time required for the AIaMD to return results. It may also risk that clinicians will cease to make their own assessment and just look at the automated assessment of segmentation

precision. The impact of such an additional feature would require evaluation before it was incorporated into the AIaMD.

7.6.3.2 Clinician actions

If adversarial testing identifies lower performance in some patient groups, such as individuals with high myopia or cataract, then clinicians should be mindful to establish patients' status with respect to these characteristics and exclude them from AIaMD use appropriately. There are three other skills that would be important for users to acquire. For photographers, a practical sense of what level and type of OCT imaging imperfections introduce a significant risk of segmentation error and warrant repeat imaging or exclusion from AIaMD analysis. For clinicians consulting with patients, it is important they understand the mechanism and performance of the AIaMD so they can effectively explain the intervention to patients, answer questions and facilitate informed consent. For clinicians making treatment decisions, it is also important that training includes why and how segmentations are inspected to decide when the AIaMD assessment of disease activity should be actioned or ignored. In our test set, CLC provided true negatives in 91.7% of the cases of major segmentation error precipitating 45.3% of R10 FPs. This shows a valuable opportunity for clinician monitoring of segmentation quality to improve the overall performance of the treatment pathway beyond the performance of clinicians or AI independently.

7.7 Discussion

Consideration of a full potential AI-enabled healthcare pathway has been facilitated by the MAA, spanning the explanation given to patients at their diagnosis of nAMD to the appointment letters they receive to maintain follow-up (see appendix). The value of this is well demonstrated as the clinical task which was attributed the highest risk score (OCT capture) does not itself involve the AIaMD at all and could easily fall outside isolated analysis of the intervention. Some of the other highest risk scores across the pathway are also attributed by interpersonal or pathophysiological mechanisms quite separate from the AI.

Beyond this more holistic observation of the AI-enabled healthcare pathway, the MAA has also led to actionable recommendations which can help to mitigate against risks and improve clinical outcomes. This includes technical actions, i.e., the thresholds applied to continuous outputs of IRF and SRF, but also operational and clinical actions. These include the importance of training photographers to understand the kind of OCT imperfections which can increase the risk of AI errors and the role clinicians can play in checking the quality of AIaMD outputs and explaining the process to patients.

7.7.1 Comparison with prior work

The MAA was designed to support healthcare providers in systematically monitoring the safety of AIaMD live within a healthcare pathway, but the value demonstrated in this chapter suggests it could be useful in developing AIaMD also.[12] This has been demonstrated in a prior published application of a MAA for an AI tool in a pre-implementation stage which diagnoses fractured necks of femurs on plain radiographs.[13] Similarly in that report, certain groups of patients (with Paget's disease) were identified as being at risk of systematically poorer treatment from an AI-enabled healthcare pathway. These insights seem important to gather prior to implementation to avoid realising these

potential harms. This value of pre-implementation place-based evaluation of AIaMD is also emphasised by a growing literature around 'silent trials'.[17, 18] This approach is increasingly recommended and aims to mitigate against the misaligned incentives of AIaMD vendors and adopters and the instability of AIaMD performance in new sociotechnical contexts.[17] The risk of being misled by vendors or unanticipated drops in performance vary between AIaMD and implementation contexts, but there is clear advantage to a systematic safeguard prior to implementation.

While place-based evaluations of AIaMD before and after implementation have been highlighted as good practice in clinical AI use, the nature and frequency of monitoring episodes remains challenging to establish. There are published perspectives from practitioners that these approaches to monitoring should be determined on a case-by-case basis from by a multi-disciplinary team containing operational, technical and clinical expertise.[18] In some academic medical centres this has been facilitated by bespoke organisational structure and committees which bring these teams together and channel any clinical AI endeavours toward their attention. Interestingly, despite their lack of communication and collaboration, very similar organisational structures seem to have developed organically with similar titles; AI centre of excellence, Clinical Intelligence Committee etc. One example from the UK is University Hospitals Birmingham NHS Foundation Trust, with a 'Digital Transformation Team' with an organisation-wide remit. In the absence of such organisational structures, engaging multiple stakeholders at an organisation remains an important step in defining what should be monitored.

The potential for MAA outputs to inform training needs for practitioners also seems highly noteworthy. This is because of the competing demands upon healthcare professionals' time, but also because of the high level of abstraction which clinical AI training guidance and literature has acted so far.[19] A recent report from Health Education England and the AI Lab acknowledged that beyond broad foundations in digital health and some AIaMD-specific concerns, clinicians main training needs would be product specific.[20] In the present case the MAA seems to provide a clear rationale for AIaMD learning outcomes for clinicians in different roles in the healthcare pathway. This seems likely to support the success of such training efforts as the need and value can be made clear to the clinicians receiving the training and less impactful learning outcomes could be avoided.

Another approach to more rigorously surface mitigations to address the risks identified by the MAA may be to conduct more in-depth explorations of clinical steps assigned a high relative risk score (Table 36). The Bow-Tie analysis, used commonly in aviation and other industries could facilitate this. Here, a particular risk is singled out, e.g. clinicians acting on a FN, and a deeper multi-disciplinary exploration of all the potential contributing mechanisms and means of mitigation is performed. The consequences of the risks and means of diminishing each of those are also explored.[21] A MAA enhanced with several Bow-Tie analyses of important risk may also offer greater insight into how workforce training could improve outcomes in an AI-enabled healthcare pathway.

7.7.2 Limitations

A major limitation is the indirect or absent engagement from stakeholders in the various stages of the MAA. The process was informed by the data collected through the qualitative research interviews described in chapter 3, but the topic guides and their use did not target

the MAA. At the present developmental stage of this AIaMD, this seems proportionate, but if the process it to be repeated nearer the time of real-world implementation place-based stakeholders should be identified and engaged to improve the rigor of risk identification and characterisation.

The numerous rule sets trialled with segmentation outputs, risk giving an inflated sense of the potential diagnostic accuracy of the tool by over-fitting to the single dataset used. The apparent advantage of R10 over the other 11 rule sets tested through this chapter and chapter 5 requires validation in other distinct datasets. Given the varying practices and perspectives of ophthalmologists across the country however, it is likely to be advantageous for implementation if an AIaMD does not dictate a given rule set with which clinicians must interpret trends in IRF or SRF. Conversely, adoption by clinicians may be enhanced if an AIaMD permits service leads to establish rule sets which make sense to them within their local instance of the AIaMD. It may also be that such rule sets are better held within departmental protocols (as they are in conventional approaches to care) rather than coded into the AIaMD, where they will incur additional regulatory demands on manufacturers.

As highlighted in the results section, adversarial testing was not completed but appears to be important to inform end-user training or contraindications to AIaMD use. Examination of major segmentation errors seems to suggest that the tool may perform less well on patients with myopic fundi, opacification of ocular media or large central pigment epithelial detachments. This would best be performed on real patient imaging selected for the presence of these features, but could also be facilitated at larger scale by synthetic images produced by a diffusion model or generative adversarial network trained to output scans with these characteristics.[22]

The test dataset was randomly selected from clinical visits of patients with nAMD made in a Newcastle clinic, which also serves Northumberland. As a result, there is almost no ethnic diversity within the dataset as census data in Newcastle suggest that 95.3% of individuals over the age of 65 are white, whilst 99.3% are white in Northumberland.[23] Given that the prevalence of nAMD among Asian and African ethnic groups is reported to be 30% and 24% of that among white ethnic groups respectively, it is not surprising that just 1 of the 262 clinic visits came from an individual with a recorded ethnicity other than white.[24] Whilst there is no mechanistic reason to expect differential performance of the AIaMD across ethnic groups, the subgroup testing in the present MAA was extremely limited in this regard which was a concern of the lay Study Reference Group.

The risk scores attributed with the failure mode and effect analysis of MAA are intended to be approximate and relative measures, but nevertheless there are opportunities to improve beyond the present chapter. Firstly, discreet choice experiments with patients and other stakeholders could help to inform how different risks should be weighted comparatively. Secondly, empirical observations of each risk's frequency and detection could be made. This would help to make the relative risk scores more rigorous to prioritise monitoring resources, but also support an empirical derivation of the ideal frequency at which each risk should be monitored for.[25] Whilst it is likely that these ideal monitoring frequencies would be prohibitively costly, actionable monitoring processes could be established with a clear sense of what compromises were being made between cost and safety.

7.7.3 Future directions

To build a firmer empirical basis for contraindications to clinical use or training materials for end-users it would be beneficial to complete the adversarial testing step which has been omitted from this iteration of the MAA. This would require the curation of a dataset enriched for patients with large central pigment epithelial detachments, myopic fundi and ocular media opacification and an evaluation of the potential AIaMD's performance. To ensure the insights that this and the rest of the MAA has granted are actionable, amenable opportunities for risk mitigation should be captured within a set of learning outcomes and training materials for end-users. This need-based approaching to training material design aligns well with the tripartite AI education framework laid out recently by NHS Health Education England and the AI lab; educational groundwork, foundational and advanced AI education and Product-specific training.[20]

Evaluation of the effectiveness of the proposed mitigation for major segmentation error through photographer and clinician training is also required. This is because the suggestion that 91.7% of major segmentation errors could be identified by suitably trained clinicians assumes that all such errors would be identified and that there are not TP and TN decisions from R10 that did not also involve inconsequential segmentation errors. It would also be valuable to understand the time impact for clinicians in requiring them to undertake such a quality assurance step.

The next step in evaluating the proposed AIaMD is to test it prospectively. Such evaluations are costly and to ensure such investments are efficient it would first be best to establish if any AIaMD with the necessary functionality have already been granted market access by a relevant regulator. A systematic scoping review of regulatory databases will be required to establish this robustly as there are no reliable or accessible registries of such candidate AIaMD.[26] Prior to the conduct of such prospective evaluation, it will also be valuable to conduct a deeper analysis of the risks and potential mitigations of risks that have been identified as more serious; i.e. suboptimal quality OCT acquisition and FPs from the AIaMD. This could be delivered by a Bow-Tie analysis, or similar method, which would help to inform risk mitigation strategies and secondary outcome measures which should be collected in a prospective evaluation.[21] Such outcome measures would also facilitate empirically derived frequencies at which different safety monitoring episodes should occur. This process would rely on principles from control engineering (Shannon-Nyquist theory), dictating that any impactful monitoring system must sample a process at least twice as frequently as the rate of occurrence of the error of interest and pragmatically ten times as frequently.[25] Where compromise is necessitated between safety and cost it would also allow decision makers to make those compromises in an informed way, minimising the number of assumptions that are required and permitting trust-promoting transparency with stakeholders.

7.8 Conclusions

Across the proposed AI-enabled clinical pathway the most important risks appear to be the capture of suboptimal quality OCT imaging and the AIaMD falsely advising of increasing nAMD disease activity. This is possible to mitigate against at the technology, product and intervention level with retraining of the AI model on a training set enriched for poor quality OCTs and targeted clinical phenotypes, adjustment to the rule-based logic applied to AI-enabled OCT segmentation and training for end-users. With these mitigations it seems likely

that the overall performance of the AI-enabled intervention will be superior to either clinicians or the AIaMD acting in isolation. This superiority is in regard to the risk of both under-treatment and over-treatment, meaning that both the quality and cost of healthcare provision could be expected to improve. Prospective evaluation of this AI-enabled healthcare pathway will be required to test this assumption, but these data appear to justify the investment required.

7.9 Appendix

7.9.1 Failure Modes and Effects Analysis – Risk Mapping

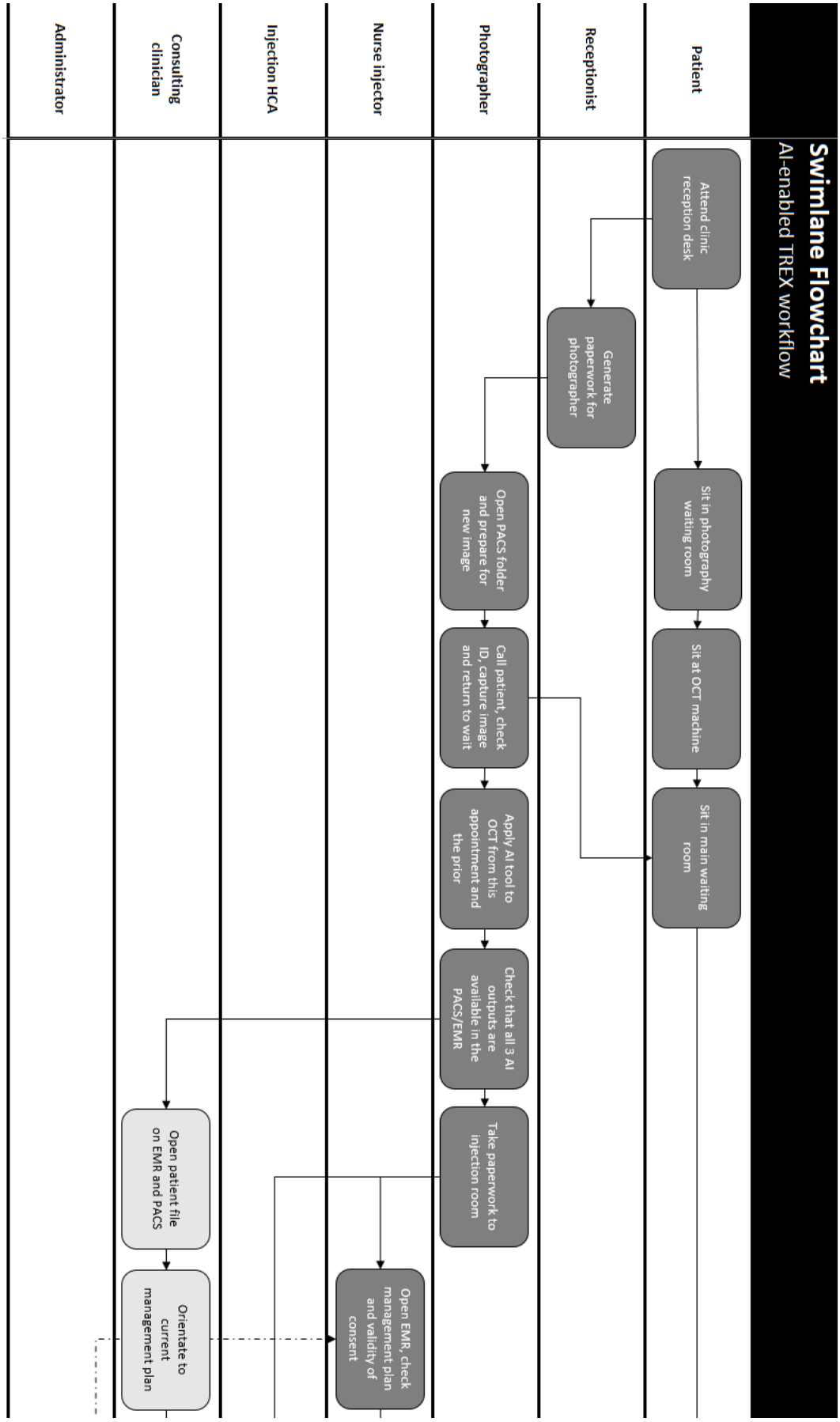
Clinical task	Elements of clinical task	Role of AI system	Risks	Risk score (severity*, occurrence**, detection***)	Mitigation strategies	Recommendation
OCT capture	Accurate ID and date	PACS user interface	Image attributed to incorrect patient	5(3+1+1)	ID check as part of imaging SOP	Design photographer training intervention
	Image quality check		Image attributed to wrong date	5(3+1+1)	Daily checking of date/time on equipment	
			Image quality increases segmentation error risk	10(3+4+3)	Photography training on image quality for AI	
Patient transfer from photography to injection	Patient explanation of pathway	Nil	Patient leaves clinic prior to receiving treatment	3(1+1+1)	Injection team to repeat calls for patient and hand over to clinicians to review images and assign urgent injection appointment	In clinician training emphasise the importance of patient explanation of the pathway
Selection and egress of OCT	Selecting image pair	PACS user interface	AI outputs based on imaging not relevent for clinical decision	5(3+1+1)	Date and laterality checks as part of photographer SOP and clinician review	Emphasis on mitigations in training programme
AI analysis of OCT	Segmentation of IRF and SRF	OCTane analysis of both OCT volume scans and output of segmentations and scalar IRF and SRF volumes	Inaccurate judgement of increasing IRF or SRF	9(2+4+3)	Clinician review of segmentation visualisations and scalar outputs in decision making	Emphasis on mitigations in clinician training programme
			Inaccurate judgement of decreasing/stable IRF or SRF	6(3+2+1)		
	Rule-based decision tree	Categorisation of disease activity or stability	Evidence of nAMD disease activity not captured by IRF and SRF thresholds of 1,000,000um3	7(4+2+1)		
Documentation of AI outputs	Recording and storing AI outputs	3 AI outputs (visualisation of segmentations, scalar IRF and SRF volumes and disease activity status) are recorded in the PACS to be clearly	AI outputs are not available for clinical decision	4(1+2+1)	Photographer SOP to include checking availability of AI output in PACS	Explore interoperability with EMR through PACS
Preparing for the injection	ID, management plan and consent checks	The management plan will be based on AI-supported clinical decisions	A patient receives an unintended injection	5(2+2+1)	Ophthalmologist availability for injectors to discuss ambiguous documentation	No change from current pathway
					ID and consent checks as part of injection	
Giving the injection	Injection	Nil	Post-injection endophthalmitis	7(4+2+1)	Sterile technique in SOP and emphatic safety netting with patients for emergency symptoms and contact details	No change from current pathway
Clinician review	Communicating review need	AI outputs support clinician's decision making	Clinical review not performed leading to patient loss to follow-up	7(4+2+1)	Clear workflow for review and dedicated staff member (fail safe officer) to check that all patients in service have complete reviews and appointments scheduled	Decision tree optimised with error analysis
	Accessing required information		Clinician wastes time reviewing useless AI output or trying to find poorly documented information	5(1+3+1)		
	Reviewing AI quality		Clinician directly actions poor quality AI output	6(3+2+1)	Clear training and SOP for clinicians on quality checking AI outputs	
	Documenting clinical decision		Administrator books inappropriate appointment type/time	5(3+1+1)		
Appointment booking	Administrator aware of clinician review	Nil	Patient does not get further treatment	7(4+2+1)	Fail safe officer review	Formalisation in SOP/patient leaflets
	Communication reaches patient successfully		Patient gets inappropriately timed treatment	7(4+2+1)	Patients given phone number to call if no appointment information within 4 weeks	
	Patient-initiated follow-up	Poor quality segmentation may indicated poor quality scans and other ocular pathologies, e.g. cataract	Patient does not report sight loss due to false reassurance that they are under review	6(3+2+1)	Injectors screen for patient concerns as part of pre-injection SOP	Could consider adding VA measurement into AI appointments
Screening for decline or ocular comorbidities	Face-to-face annual review				Dedicated line of contact for nAMD patient concerns	
Informed consent for AI-enabled treatment	Enabling patient understanding of management and rationale	Nil	Patients receive treatment they do not want	5(2+2+1)	Consent and capacity review as part of annual face-to-face appointment	Training programme development will be important here
			Patients decline treatment they would have wanted if they understood accurately	7(4+2+1)	Greater time allocation to face-to-face appointments	
					Information leaflets	
					Clinician training on communicating AI role	

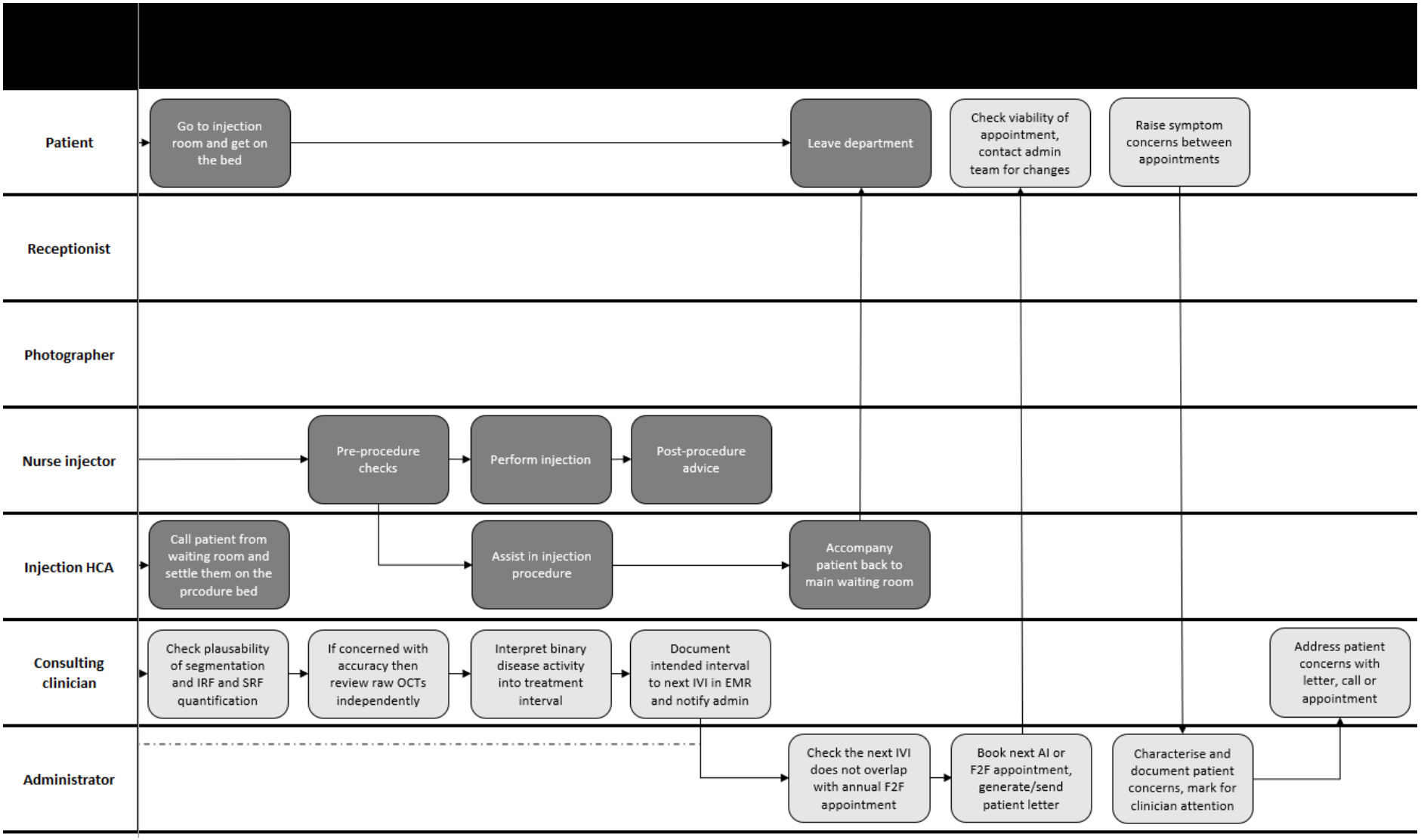
* Severity score: 4 = vision loss with one-off occurrence, 3 = vision loss with repeated occurences, 2 = >1 clinically unecessary injection/year, 1 = No harm or anxiety/inconvenience

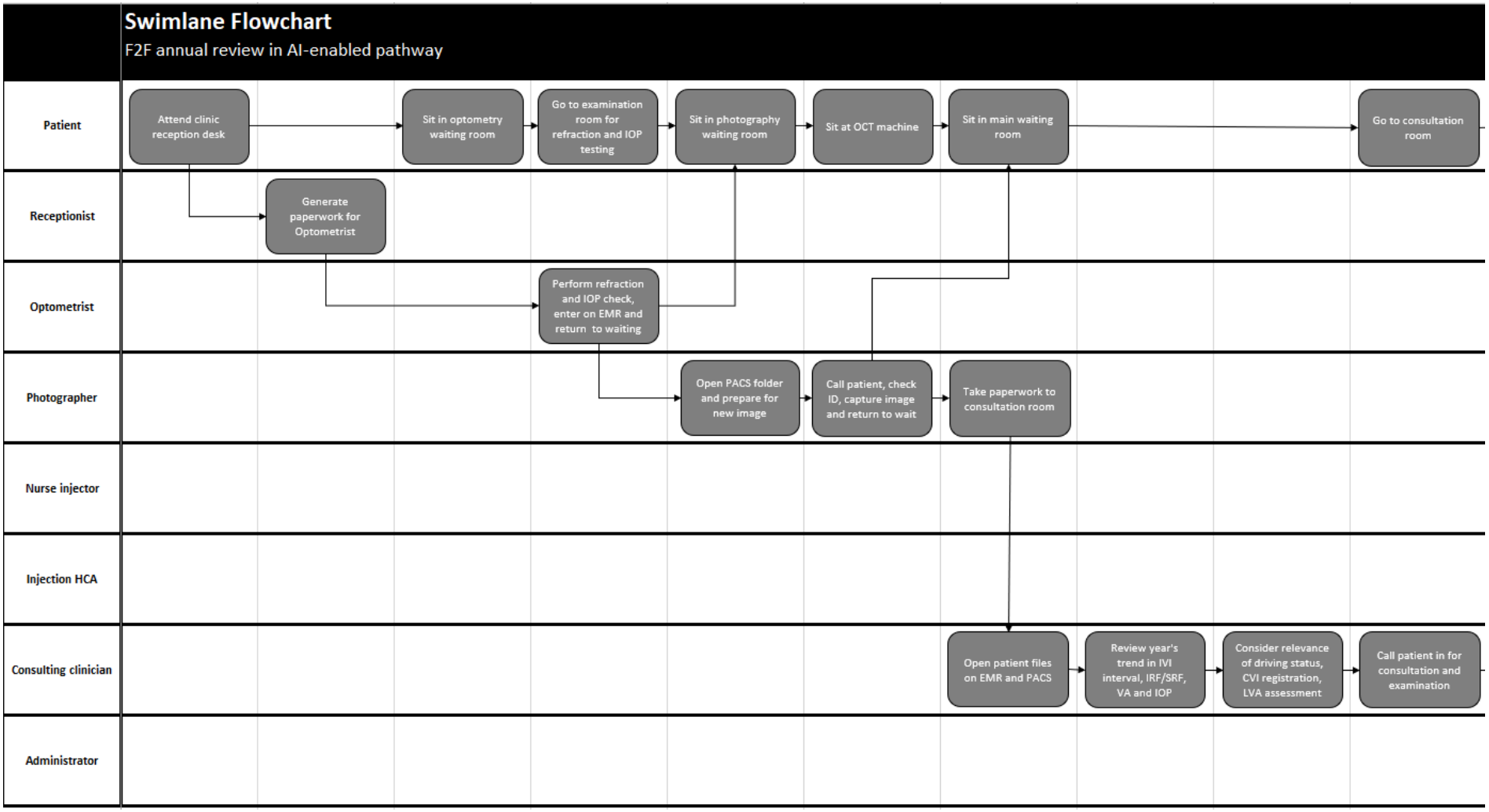
** Probability score: 4 = >10% of patient years, 3 = 1-10% of patient years, 2 = 0.1-1% of patient years, 1 = <0.1% of patient years

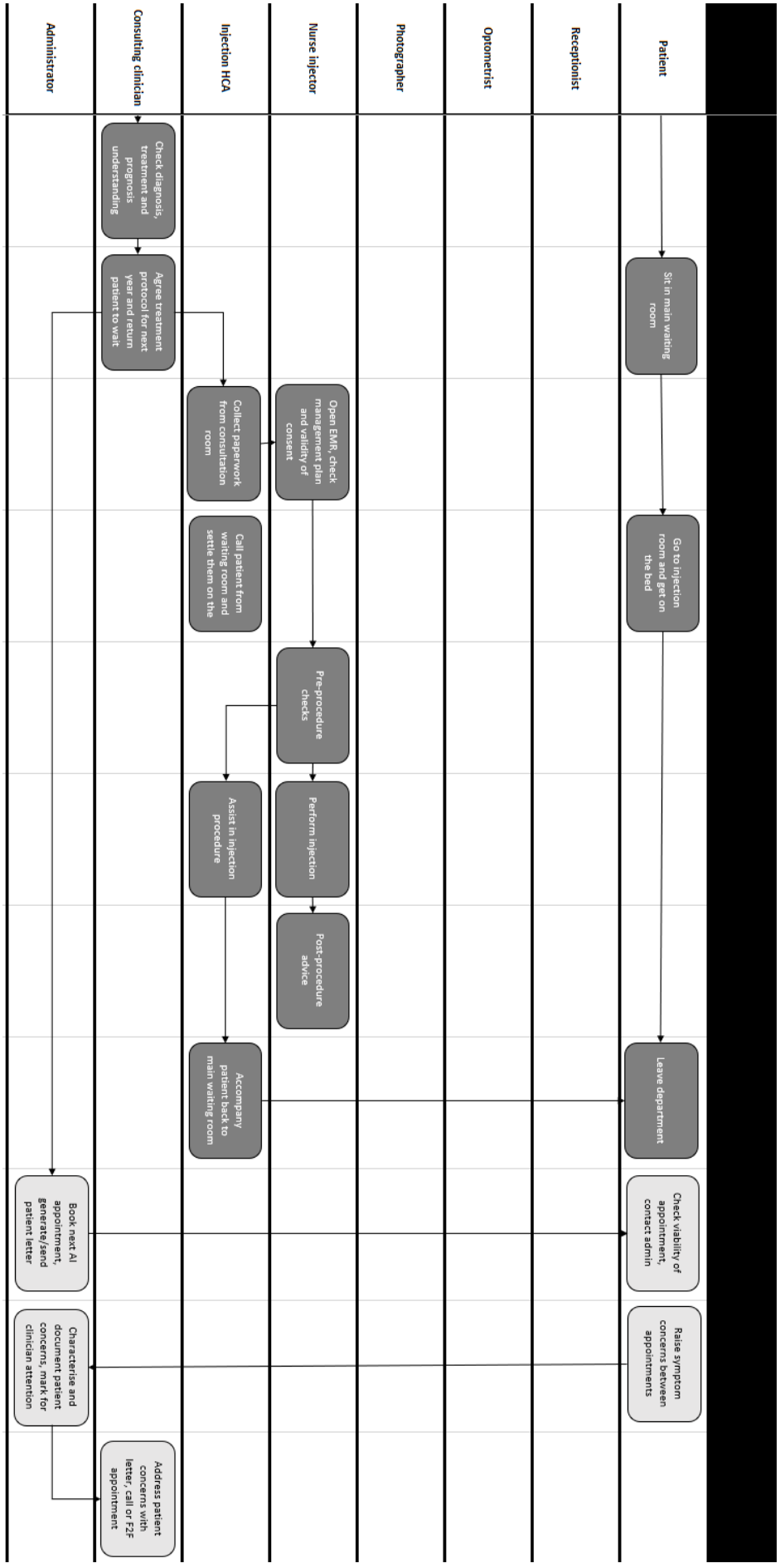
*** Detection score: 4 = <10% of errors expected to be detected, 3 = 10-50% of errors expected to be detected, 2 = >50% of errors expected to be detected, 1 = 1 or more dependable setection mechanism

7.9.2 Clinic swim lane diagrams




Patient
Receptionist
Photographer
Nurse injector
Injection HCA
Consulting clinician
Administrator
Go to injection room and get on the bed
Leave department
Check viability of appointment, contact admin team for changes
Raise symptom concerns between appointments
Pre-procedure checks
Perform injection
Post-procedure advice
Call patient from waiting room and settle them on the prcodure bed
Assist in injection procedure
Accompany patient back to main waiting room
Check plausability of segmentation and IRF and SRF quantification
If concerned with accuracy then review raw OCTs independently
Interpret binary disease activity into treatment interval
Document intended interval to next IVI in EMR and notify admin
Address patient concerns with letter, call or appointment
Check the next IVI does not overlap with annual F2F appointment
Book next AI or F2F appointment, generate/send patient letter
Characterise and document patient concerns, mark for clinician attention


Swimlane Flowchart
F2F annual review in AI-enabled pathway
Patient
Receptionist
Optometrist
Photographer
Nurse injector
Injection HCA
Consulting clinician
Administrator
Attend clinic reception desk
Sit in optometry waiting room
Go to examination room for refraction and IOP testing
Sit in photography waiting room
Sit at OCT machine
Sit in main waiting room
Go to consultation room
Generate paperwork for Optometrist
Perform refraction and IOP check, enter on EMR and return to waiting
Open PACS folder and prepare for new image
Call patient, check ID, capture image and return to wait
Take paperwork to consultation room
Open patient files on EMR and PACS
Review year's trend in IVI interval, IRF/SRF, VA and IOP
Consider relevance of driving status, CVI registration, LVA assessment
Call patient in for consultation and examination


Patient
Receptionist
Optometrist
Photographer
Nurse injector
Injection HCA
Consulting clinician
Administrator
Sit in main waiting room
Go to injection room and get on the bed
Leave department
Check viability of appointment, contact admin
Raise symptom concerns between appointments
Open EMR, check management plan and validity of consent
Collect paperwork from consultation room
Call patient from waiting room and settle them on the
Pre-procedure checks
Perform injection
Post-procedure advice
Assist in injection procedure
Accompany patient back to main waiting room
Check diagnosis, treatment and prognosis understanding
Agree treatment protocol for next year and return patient to wait
Book next AI appointment, generate/send patient letter
Characterise and document patient concerns, mark for clinician attention
Address patient concerns with letter, call or F2F appointment

7.9.3 Screening for case characteristics associated with R10 FNs

Table 40. Screening for unequal performance of OCTane outputs interpreted by rule set 10 (R10) between clinic visits with different categorical characteristics using descriptives of the absolute number of false negatives (FN) and the overall false negative rates (FNR) in between different groups. Chi-squared tests are used to derive p values.T2DM = Type 2 Diabetes Mellitus, TEX = Treat-and-Extend.

		not FN	FN	FNR	p value
Number of eyes treated	Unilateral treatment	182	3	1.6%	
	Ongoing bilateral treatment	75	2	2.6%	0.6
Sex	Female	157	4	2.5%	
	Male	100	1	1.0%	0.39
Ethnicity	British	238	5	2.1%	
	Pakistani	1	0	0.0%	
	Not stated	18	0	0.0%	0.82
Laterality	Left	118	4	3.3%	
	Right	139	1	0.7%	0.13
Drug	Aflibercept	234	4	1.7%	
	Ranibizumab	23	1	4.2%	0.4
Diabetic status	Not diabetic	150	4	2.6%	
	T2DM	44	1	2.2%	
	Diabetic - unknown type	4	0	0.0%	
	Status unknown	59	0	0.0%	0.65
Protocol	Loading	53	0	0.0%	
	TEX	204	5	2.4%	0.26
Macular haemorrhage	no	251	5	2.0%	
	yes	6	0	0.0%	0.73

Table 41. Screening for unequal performance of OCTane outputs interpreted by rule set 10 (R10) across clinic visits with different continuous characteristics using descriptives of the mean and 95% confidence intervals (CI) of visits which were and were not assigned a false negative (FN) assessment of disease activity by CLC. Independent t- tests are used to derive p

values comparing the difference of the means between visits which were and were not assigned a FN assessment. IDAOPI = Income Deprivation Affecting Older People Index, VA = Visual Acuity, nAMD = neovascular Age-relate Macular Degeneration

N=262 (unless otherwise stated		IDAOPI	Years of age	Baseline VA (letters)	Prior treatment interval	Prior injections	Visit VA (letters)	Contralateral VA (letters)	most recent nAMD activity (weeks) N = 176
not R10 FN n=257	Mean	5.2	80.4	60.2	8.7	12.7	63.0	56.8	26.0
	upper CI	5.6	81.3	61.9	9.1	13.7	64.9	60.3	29.4
	lower CI	4.9	79.4	58.6	8.4	11.6	61.2	53.4	22.7
R10 FN n=5	Mean	4.8	81.0	51.4	10.0	15.6	61.2	63.4	40.6
	upper CI	6.4	85.1	62.7	17.5	22.3	74.5	87.1	73.5
	lower CI	3.2	76.9	40.1	2.5	8.9	47.9	39.7	7.7
p value		0.62	0.78	0.20	0.76	0.44	0.80	0.62	0.44

7.9.4 Screening for case characteristics associated with R10 FPs

Table 42. Screening for unequal performance of OCTane outputs interpreted by rule set 10 (R10) between clinic visits with different categorical characteristics using descriptives of the absolute number of false positives (FP) and the overall false negative rates (FPR) in between different groups. Chi-squared tests are used to derive p values.T2DM = Type 2 Diabetes Mellitus, TEX = Treat-and-Extend.

		not FP	FP	FPR	p value
Number of eyes treated	Unilateral treatment	150	35	18.9%	0.41
	Ongoing bilateral treatment	59	18	23.4%	
Sex	Female	128	33	20.5%	0.89
	Male	81	20	19.8%	
Ethnicity	British	193	50	20.6%	0.81
	Pakistani	1	0	0.0%	

	Not stated	15	3	16.7%	
Laterality	Left	95	27	22.1%	
	Right	114	26	18.6%	0.47
Drug	Aflibercept	190	48	20.2%	
	Ranibizumab	19	5	20.8%	0.94
Diabetic status	Not diabetic	124	30	19.5%	
	T2DM	33	12	26.7%	
	Diabetic - unknown type	3	1	25.0%	
	Status unknown	49	10	16.9%	0.65
Protocol	Loading	40	13	24.5%	
	TEX	169	40	19.1%	0.38
Macular haemorrhage	no	203	53	20.7%	0.21
	yes	6	0	0.0%	

Table 43. Screening for unequal performance of OCTane outputs interpreted by rule set 10 (R10) across clinic visits with different continuous characteristics using descriptives of the mean and 95% confidence intervals (CI) of visits which were and were not assigned a false positive (FP) assessment of disease activity by CLC. Independent t- tests are used to derive p values comparing the difference of the means between visits which were and were not assigned a FP assessment. IDAOPI = Income Deprivation Affecting Older People Index, VA = Visual Acuity, nAMD = neovascular Age-relate Macular Degeneration

N=262 (unless otherwise stated		IDAOPI	Years of age	Baseline VA (letters)	Prior treatment interval	Prior injections	Visit VA (letters)	Contralateral VA (letters)	most recent nAMD activity (weeks) N = 176
not R10 n=209 FP	Mean	5.2	80.7	60.1	8.8	13.2	63.5	56.8	25.4
	upper CI	5.6	81.7	61.9	9.1	14.4	65.5	60.6	29.0
	lower CI	4.9	79.7	58.3	8.4	12.0	61.5	52.9	21.7
	Mean	5.2	79.1	59.9	8.8	10.9	61.0	57.6	32.1

R10 FP n=53	upper CI	6.0	81.4	63.8	9.6	12.8	65.4	64.7	40.7
	lower CI	4.4	76.9	55.9	8.1	9.0	56.6	50.5	23.6
p value		0.88	0.22	0.92	0.86	0.05	0.31	0.84	0.16

7.10 References

1. Hogg, H.D.J., et al., Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence. J Med Internet Res, 2023. 25: p. e39742.

2. Skivington, K., et al., A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. Bmj, 2021. 374: p. n2061.

3. Greenhalgh, T., et al., Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res, 2017. 19(11): p. e367.

4. Lyell, D., et al., More than algorithms: an analysis of safety events involving ML-enabled medical devices reported to the FDA. J Am Med Inform Assoc, 2023. 30(7): p. 1227-1236.

5. The United States Food and Drug Administration, K212783 - ProstatID approval letter. 2022.

6. The British Standards Institute, The European Medical Devices Regulations – What are the requirements for vigilance reporting and post-market surveillance? 2020.

7. The Accelerated Access Collaborative, Public perceptions and attitudes to Artificial Intelligence (AI) in healthcare. An exploratory study. 2022.

8. Al-Zubaidy, M., et al., Stakeholder Perspectives on Clinical Decision Support Tools to Inform Clinical Artificial Intelligence Implementation: Protocol for a Framework Synthesis for Qualitative Evidence. JMIR Res Protoc, 2022. 11(4): p. e33145.

9. Carayon, P., et al., Work system design for patient safety: the SEIPS model. Qual Saf Health Care, 2006. 15 Suppl 1(Suppl 1): p. i50-8.

10. Leveson, N., A new accident model for engineering safer systems. Safety Science, 2004. 42(4): p. 237-270.

11. International Electrotechnical Commission, IEC 61882 - Hazard and operability studies (HAZOP studies) – Application guide. 2016.

12. Liu, X., et al., The medical algorithmic audit. The Lancet Digital Health, 2022. 4(5): p. e384-e397.

13. Oakden-Rayner, L., et al., Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. The Lancet Digital Health, 2022. 4(5): p. e351-e358.

14. The Medicines and Healthcare products Regulatory Agency, Crafting an intended purpose in the context of Software as a Medical Device (SaMD). 2023.

15. De Fauw, J., et al., Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med, 2018. 24(9): p. 1342-1350.

16. Sendak, M.P., et al., Presenting machine learning model information to clinical end users with model facts labels. npj Digital Medicine, 2020. 3(1): p. 41.

17. Kwong, J.C.C., et al., The silent trial - the bridge between bench-to-bedside clinical AI applications. Front Digit Health, 2022. 4: p. 929508.

18. Kim, J.Y., et al., Organizational Governance of Emerging Technologies: AI Adoption in Healthcare, in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023, Association for Computing Machinery: Chicago, IL, USA. p. 1396–1417.

19. Khurana, M.P., et al., Digital health competencies in medical school education: a scoping review and Delphi method study. BMC Medical Education, 2022. 22(1): p. 129.

20. NHS AI Lab, NHS Transformation Directorate and NHS England, Developing healthcare workers' confidence in artificial intelligence (AI) (Part 2). 2022.

21. Poleto, T., et al., A Risk Assessment Framework Proposal Based on Bow-Tie Analysis for Medical Image Diagnosis Sharing within Telemedicine. Sensors (Basel), 2021. 21(7).

22. Yoo, T.K., et al, Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification. Med Biol Eng Comput, 2021. 59(2): p. 401-415.

23. Office for National Statistics, Population estimates. 2021.

24. Zhou, M., et al., Geographic distributions of age-related macular degeneration incidence: a systematic review and meta-analysis. Br J Ophthalmol, 2021. 105(10): p. 1427-1434.

25. Mohtadi, C., Bode's integral theorem for discrete-time systems. IEE Proceedings D (Control Theory and Applications), 1990. 137(2): p. 57-66.

26. Ong, A., et al., Artificial Intelligence as a Medical Device for Ophthalmic Imaging in Europe, Australia, and the United States: Protocol for a Systematic Scoping Review of Regulated Devices. Open Science Framework, 2023.

Chapter 8: Thesis discussion

This chapter begins by recalling the thesis' stated aim and summarising the findings of chapters 3 – 7 in its pursuit. These findings are then interpreted together to answer the central questions of what factors sustain the AI chasm generally, and in the specific case of AI-enabled macula services. The chapter concludes with actionable solutions for the implementation of AI-enabled macula services, recommendations for researchers, recommendations for practitioners and concluding remarks for the thesis.

8.1 Summary of findings

This thesis aimed to explore the factors which sustain the AI chasm in healthcare generally and develop actionable solutions for the implementation of AI-enabled macula services in the NHS. This was achieved through a programme of research which established suitable methods for qualitative research of clinical AI generally and then applied them within the target context to evaluate a potential AI technology, the AIaMD in which it might be embedded, the intervention in which that AIaMD might be embedded, and finally the healthcare pathway in which that intervention might be embedded (Figure 42). Relevant publications and presentations made at the time of writing are referenced in the passage below.

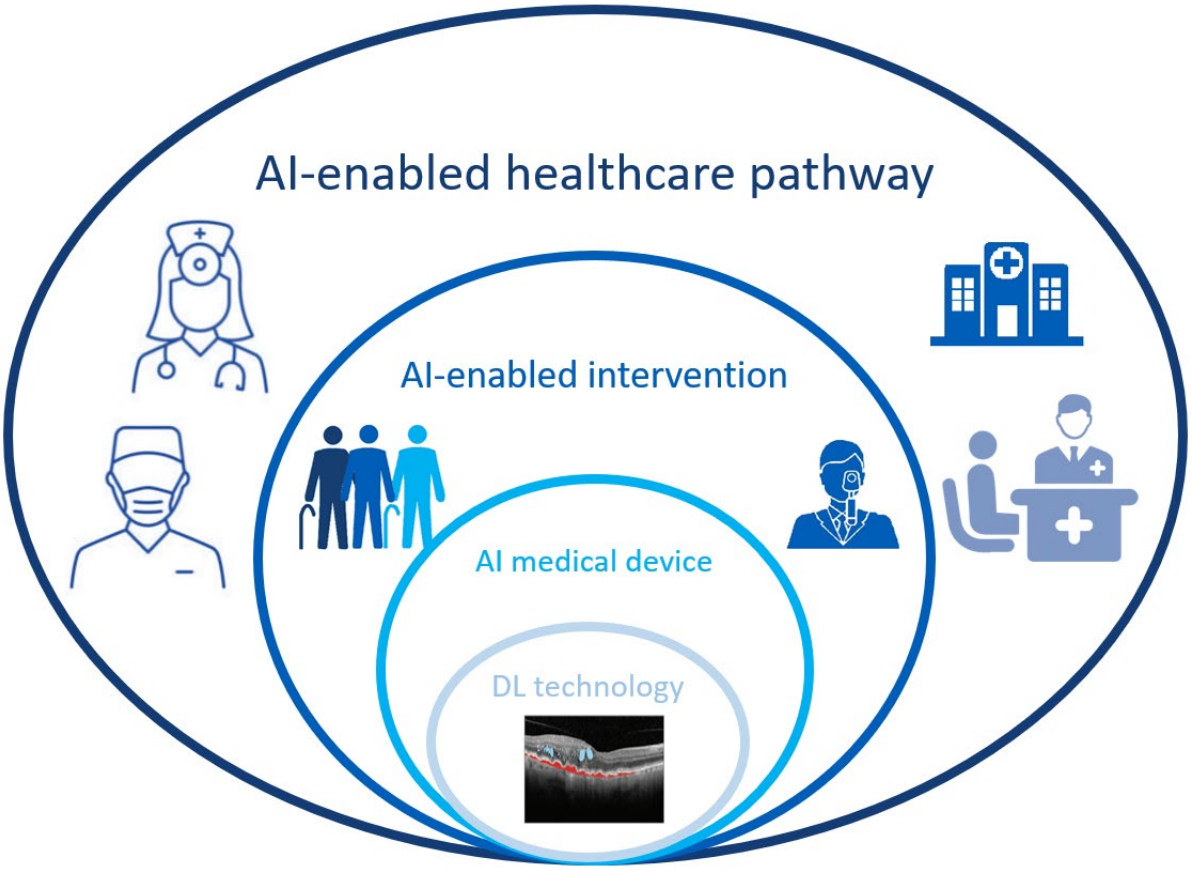


Figure 42. Schematic of the relationships between a deep learning (DL) technology such as OCTane, the artificial intelligence (AI) medical device in which it sits, the AI-enabled intervention in which that sits and the AI-enabled healthcare pathway in which the intervention sits.

The qualitative evidence synthesis of chapter 3 helped to characterise 5 highly abstracted stakeholder groups in clinical AI implementation.[1, 2] Each of these groups showed the potential to influence implementation outcomes but tended to have distinct means and motives to do so. It was also noteworthy that these stakeholders, and the factors that they highlighted, exert their influence on implementation outcomes in an interdependent manner. This means that the successful implementation of any given AI-enabled intervention can be expected to depend on many stakeholders and interrelated factors.[3] Unfortunately for practitioners seeking to implement clinical AI, insights into the perspectives of all these stakeholder groups, besides clinicians, are scarce in the literature.[4, 5] Another challenge for clinical AI implementation researchers, in making

sense of the data they elicit, is the variation of TMFs applied across the literature, often selected without an explicit rationale. Chapter 3 helped to address these limitations by consolidating factors that implementation practitioners and researchers can consider exploiting in their own work and curating a list of TMFs previously used in clinical AI research to choose from.[6]

These foundations, intended to be relevant across clinical AI implementation, were used to design the method applied in chapter 4.[7] They informed the stakeholder groups recruited, the topic guides used to collect data and the TMF taken to data analysis. A key finding was that stakeholders in nAMD generally welcome the prospect of AI technology, with a broad potential value proposition envisaged including service efficiency, patient and clinician experience, care consistency, clinical outcomes, care transparency, equity of access and environmental impact.[5] However, consultant ophthalmologist perspectives are likely to be particularly influential over implementation outcomes and an AI-enabled macula service would not be accepted if it compromised patients' visual outcomes.

Chapter 5 took this minimum requirement from stakeholders as the non-inferiority endpoint for an observational study of the NPV of AI-led assessments of nAMD activity or stability compared to real-world clinical assessments.[7] As the AI technology evaluated does not provide this necessary simple binary output, different AIaMDs were simulated by applying clinically inspired rule-sets to the OCT segmentation outputs that the technology generates. The sample size for the non-inferiority test was calculated from a pilot data set and a simple rule that any increase in IRF between sequential clinic visits signalled nAMD activity was applied to AI segmentations of OCTs. This simple initial AIaMD design to blend deep learning and rule-based AI components clearly satisfied the clinical and statistical thresholds for non-inferiority with a rNPV of 1.06 and a 95% CI of 0.99 – 1.13 (rPPV 1.05, 95% CI 0.84 – 1.32). Exploration of a few additional clinically intuitive rule sets helped to improve on this performance with a more complex definition of nAMD activity being signalled by a 10% or higher increase in IRF or SRF producing a statistically superior rNPV of 1.18 and a 95% CI of 1.09 – 1.27 (rPPV 1.00, 95% CI 0.82 – 1.23).[8]

This established that existent AI technology could be situated within an AIaMD and be expected to improve clinical decisions rather than maintain or worsen them. The focus of the thesis then shifted to the intervention in which the AIaMD should be embedded. Chapter 6 aimed to align this AI-enabled intervention with stakeholder values and the factors likely to influence implementation. This practical goal was supported by the data elicited in chapter 4 and a choice of TMF for their analysis that focused on digital health innovation and accommodated factors at the individual, organisational and policy levels. This TMF was selected from the systematically curated list of TMFs in chapter 3. Despite future opportunities to expand value with greater autonomy of the AIaMD and decentralisation of nAMD care pathways an initial intervention was proposed to balance the risk of abandonment against the value proposition. This entailed an AIaMD acting on OCT data hosted on a cloud platform, interoperable with NHS provider's PACS and EMR. A decision support application of the AIaMD on treatment monitoring episodes interspersed within annual F2F consultations for patients with a known nAMD diagnosis was recommended. Task-shifting of nAMD monitoring consultations away from consultant ophthalmologists, with greater involvement of more junior nursing staff in injection delivery and photographers in the operational facets of AIaMD application was also

recommended.[5] Given that the balance of different stakeholder and factor influence over implementation outcomes will vary between settings and over time, another key characteristic of this analysis was its transparency. This transparency ought to enable future re-iterations of the proposed intervention to tailor it to specific implementation efforts.[9]

With a clearer sense of the AI-enabled intervention in mind it became possible to specify an IUS for a hypothetical SaMD that developers could aim to achieve regulatory approval for. It also became possible to project the care pathway in which the intervention would be delivered, to anticipate risks and their mitigation, using a MAA approach in chapter 7. The insights gained suggested that further revision of the rule-based component of the AIaMD could improve performance further. Taking increases in IRF over 1,000,000 μm^3 or increases in SRF over 2,000,000 μm^3 to signal nAMD activity and comparing to CLC decisions from chapter 5 estimated rNPV at 1.19 with a 95% CI of 1.11 – 1.28 (rPPV 1.31, 95% CI 1.06 – 1.60). These statistics suggest a wider potential value proposition for AI-enabled macula services, with a statistically significant lower rate of wasteful over-treatment (described by rPPV) and statistically and clinically significant superiority for avoiding sight-threatening under-treatment (described by rNPV). Recommendations for further improvements on the outcomes of AI-enabled macula services were also made. For developers, attempting to improve the performance of the deep learning technology on myopic fundi or those with large pigment epithelial detachments would be beneficial. Means of integrating some form of output uncertainty indicator were also proposed. For developers, roles for clinicians acting as human-in-the-loop were proposed that maintained the efficiency and task-shifting value proposition of AI-enabled macula services whilst improving safety and effectiveness. Learning outcomes for photographers and clinicians were also proposed for user training.

8.2 Interpretation of findings

This thesis studies a clinical AI technology for which a rigorous external validation was published in a high impact academic journal (Nature Medicine) more than 5 years ago.[10] The factors that have supported the persistence of the "AI chasm" for this technology are clearly not purely technological.[10, 11] Instead, they mainly relate to the breadth of stakeholder groups involved in the implementation of clinical AI and the social factors that limit their mutual understanding and co-operation.[1] Contemporaneous examples of this are manifold:

- policy makers restrict AI investment to specific clinical specialties [12]
- regulators have not delivered transparency over the intended use and observed risks of SaMD [13, 14]
- developers prioritise global markets besides the NHS for their greater return on investment [15]
- managers prioritise scaling up familiar approaches to create capacity gains [16]

This challenge is not unique in healthcare innovation, even within the more specific scope of digital health of CDS.[17, 18] It may however be more pronounced within clinical AI, by the widely perceived novelty and complexity of the technical, operational and clinical considerations it requires.[1, 19] These challenges must be accommodated within any plans to implement AI-enabled macula services. This thesis' findings should inform and motivate, rather than discourage, efforts to overcome the challenges which sustain this instance of the "AI chasm".

A systematic review of 4 global regulatory databases has already highlighted 2 approved SaMD (since 2022) that would appear to facilitate the decision support on which an NHS AI-enabled macula service could be based.[20] This as-yet unpublished systematic review also highlighted that the only other regulated clinical use cases for AI-enabled ophthalmic imaging analysis were diabetic retinopathy screening, optic neuropathy screening and cardiovascular risk prediction. There is no precedent for the implementation of the 2 AIaMD relevant to this thesis within the NHS at the time of writing. However, several NHS trusts considering AI-enabled macula services are in pre-procurement local validation stages with one or both AIaMD. Some stakeholders will find it concerning that these implementation decisions are being entered into without a rigorous prospective evaluation or formal health technology assessment from NICE. However, this is becoming established as the norm as NICE has approved only a fraction of the AIaMD on market and hundreds of NHS trusts use AIaMD that NICE has explicitly said it cannot yet endorse.[21, 22] An expectancy of NICE approval for AIaMD has not been routinely held for physical medical devices adopted into practice. However, limited or misaligned understanding of AI technologies and what their closest established comparator is has led many to expect levels of evidence and assurance associated with the implementation of new drugs.[23] The observational, low-resource and pragmatic methods and findings of this thesis could form a template for setting-specific evidence generation to inform local decision makers in the absence of NICE approvals. Such place-based evaluations are central to guidelines on the practice of AI implementation, but also on the scientific requirement for the replicability of research findings.[24, 25] They could also help to identify and meet local stakeholders' needs for evidence that closely relates to their experience of nAMD care and may still be able to contribute to NICE's new Early Value Assessment approach for digital health innovations.[5, 26]

In designing the AI-enabled intervention for nAMD treatment monitoring, an important question over the specific contribution that AI makes to the value proposition. This is because much of the type and scale of value proposed could be achieved through a highly centralised telemedical model of care for nAMD monitoring.[27, 28] The closest example to this potential centralised telemedical monitoring service is the long-established NHS diabetic retinopathy screening service. Notably, this service continues under sustained pressure to improve efficiency through AI-enablement, despite several failed attempts to reach the threshold of evidence demanded by the relevant decision makers.[29] Taken alongside the present thesis, this observation suggests a perception by senior leadership that AI-enabled models of care can offer greater value than telemedical models of care within the NHS, even when the evidence base is not yet adequate to inform implementation. Perhaps even more pertinent than the potential impotence of AI to meaningfully exceed the value proposition of telemedical nAMD treatment monitoring, is the question of why examples of telemedical models of nAMD monitoring have failed to spread. There are established telemedical models of care at some NHS trusts (e.g. University Hospitals Birmingham and York and Scarborough Teaching Hospitals NHS Foundation Trusts) whilst others (e.g. NuTH NHS Foundation Trust) continue with F2F consultations distinct from treatment administration (usually at the same hospital visit). This suggests that the value proposition of AI-enabled macula services may in practice be inadequate to motivate early adoption, or that its value proposition will need to prove more compelling than that of a telemedical alternative (perhaps through decision automation). For macula services yet to adopt either telemedical or AI-enabled models of care, the current national policy backdrop

set on driving clinical AI innovation may limit the need for AI cost-effectiveness to motivate instances of early adoption.[12] In the absence of interventional evidence of clinical and cost-effectiveness, sustainment of AI-enabled macula services beyond the current policy context, will be less certain. A significant mitigation for NHS services is that much of the digital infrastructure required for AI-enabled macula services is also needed for telemedical services. Even if investments in AI-enabled macula services prove to be misplaced then, they will also enable telemedical services as a contingency.

8.3 Residual barriers to AI-enabled macula services

8.3.1 Regulatory barriers

At present there are no SaMD regulated for autonomous clinical decisions based on OCT analysis.[20] Class IIa, under which the CE marks were attributed to the two SaMD described above, accommodates SaMD to be used to 'inform' or 'drive' clinical management in 'serious' healthcare conditions such as that posed by nAMD.[30] The precise AI-enabled workflows which can be conducted in line with these descriptions is open to interpretation, but are likely to include some form of viable AI-enabled macula service. In its current state however, the responsibility for interpreting this ambiguity will lie with clinical risk teams in adopting organisations for whom clinical AI most commonly represents unfamiliar territory. Unless the risk appetite and competencies of these teams can accommodate the ambiguity of the current available IUS, even low-autonomy AI-enabled macula services will fail to be adopted. Even where adoption proves successful, restrictions on the autonomy of the SaMD may also limit the value proposition that the hosting AI-enabled macula service can offer stakeholders. As such, it would be advantageous if regulatory approval could be sought for a class IIb use case to 'treat or diagnose' nAMD, either by the manufacturer of one of these existent SaMDs or an additional manufacturer.[30] Initiating evidence generation for this is particularly urgent given that 71% of SaMD regulatory submissions to European Notified Body (NB) took 13 months or longer from submission to certificate and market access in a recent survey.[31]

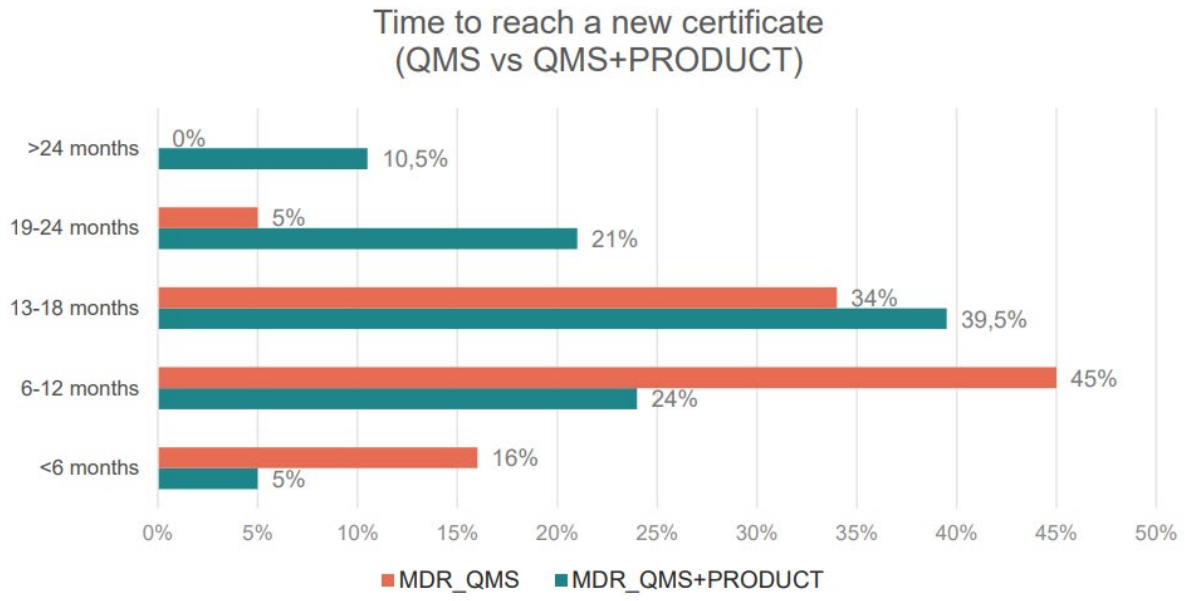


Figure 43. Histogram showing time to certification from submission for quality management systems (QMS) or QMS and medical device products among European notified bodies (NB). Reproduced from a 2023 survey of 39 NBs.[31]

8.3.2 Operational barriers

A further limit to the value proposition and hence implementation of AI-enabled macula services comes from the current dependency of NHS services on centralised injection facilities. The nature of nAMD treatment dictates that patients regularly receive injections.[32] If this must be on hospital sites then there is little incentive to re-distribute imaging and decision-making tasks to the community, preventing the fuller value proposition of decentralised nAMD care. This dependence on centralised injection administration could be partially mitigated by achieving the service capacity required by PRN treatment regimens. If PRN regimens become more prevalent across the NHS, a reduction in the proportion of AMD patients actively receiving injections might be expected.[33, 34] This could enable greater task-shifting to community optometry led monitoring of quiescent nAMD, in turn lowering the burden on hospital eye services.[35] A greater impact would be made if the infrastructure required for injections could be shifted to the community, either through purpose-built community satellite clinics for NHS trusts, hosting of nAMD treatment clinics in established primary care premises or upscaling of mobile injection room initiatives.[27]

8.3.3 Evidence limitations

The evidence generated within this thesis has clear limitations which further research can improve on. Seeking to replicate the observational non-inferiority studies with data sets from distinct NHS services is the first requirement. This will help to check the robustness of the finding of superior NPV and PPV from this work and the risks and mitigations proposed through the algorithmic audit.[25, 36] Similarly, qualitative data collection should also be extended across diverse NHS sites to test the validity of the assumptions that underly the proposed AI-enabled intervention design and seek to broaden the scope of its relevance. Qualitative findings were not just biased by their narrow geographical focus, but also their means of collection and analysis. As a trainee ophthalmologist and colleague of many participants, the lead researcher (JH) brought their own personal bias to study design, data collection and data analysis. The interviewer made no effort to actively disclose their clinical role to participants, but it was often asked about in conversation and all questions were answered honestly. To allow readers to critically appraise the work, this bias has been candidly reported and mitigation was sought through input from the study reference and advisory groups and reflective journalling after each of the interviews. One such example was the suggestion from the lead researcher and supervisors that recruitment was complete, which was then challenged by a member of the reference group, leading to the recruitment of a further charity sector representative.

Resource preservation is the key value proposition for many stakeholders and so the absence of methods, e.g. AI-enabled workflow simulation, to measure resource impact is another important limitation. Along with other data, simulated workflow observations can facilitate an economic evaluation to compare AI-enabled macula services to contemporaneous models of nAMD care in the NHS. Assuming these investigations do not refute the value proposition of AI-enabled macula services, an interventional evaluation of one, but preferably several, AIaMD should be conducted. This evaluation should pivot from the diagnostic accuracy outcomes used in this thesis' observational work to real-world outcome measures such as VA, injection frequency and adverse event reporting.[23, 37] To maximise its utility, such an evaluation should ideally collect additional data seeking to test

and validate the wider value proposition explored in this thesis; patient and clinician experience, service equity and environmental impact.

8.4 Recommendations

Research is the systematic process used to generate new knowledge or evidence if AI-enabled macula services are to be accepted by decision makers at local, regional or national levels. Though this is often attributed to 'researchers', e.g. university academics and industry professionals, 'practitioners', e.g. NHS clinicians, managers and technologists, produce their own forms of evidence to inform a range of decisions though typically at a local scale. To maximise the productivity and synergy of both, it is important to consider the distinction between researchers and practitioners and their relative strengths in generating evidence. Besides their role in evidence generation, practitioners also play an important role of interpreting evidence in the context of their practice and executing innovation.

8.4.1 Recommendations for researchers

Research is constrained by the long timescales of applying for funding, rigorous processes in conducting the research and peer-review publication. It often aims to generate knowledge because of its intellectual or societal value, rather than short term commercial or practical value. This confers a distinct advantage in the scope of aims which it can attract investment to, allowing for the investigation of problems experienced by stakeholders that may otherwise be overlooked of solutions offering less immediate or guaranteed value back. The systematic methods and reporting standards expected of researchers can also lend the evidence they generate a high degree of credibility and so influence decision makers with large jurisdictions. The MHRA or NICE are examples of such decision makers who have the potential to define or even obligate the use of AIaMD across the NHS.

Within the setting of the present thesis and current efforts by NHS adopters to implement AI-enabled macula services, it seems that motivating early adoption in the NHS is no longer a valuable target for new research. The time constraints described above suggest that evidence that researchers begin planning the generation of now will come to fruition in the midterm (e.g. 2027-2030). At this time it seems likely that testing the relative benefits of the AI-enabled macula services that early adopters have developed and their suitability for scaling and spread across the NHS will be of greatest value.[23] Consequently, researchers should prioritise generating high quality evidence of safety, clinical and cost effectiveness for the uses of AIaMD in nAMD treatment monitoring that emerge from practitioners' early-adoption experiences. This evidence should aim to support decisions by entities such as NICE and the MHRA which will enable the scaling of these models of AI-enabled nAMD care. Researchers should also exercise their freedom to pursue non-financial value to refine and evaluate interventions that maximise improvements in AMD service equity, environmental impact, user training and safety monitoring processes. Beyond the adoption of these initial interventions, researchers can then lift their focus from hybrid effectiveness-implementation research to focus on the evaluation and refinement of implementation strategies rather than the intervention itself.[38] This may employ more traditional approaches which aim to identify the 'right' approach, e.g. randomised controlled trials of implementation strategies, or approaches that embrace complexity and aim to answer what works for who, why and under what circumstances, e.g. critical realism.[39, 40]

8.4.2 Recommendations for practitioners

The implementation of AI-enabled macula services, like clinical AI more generally, will depend upon varied teams with a breadth of clinical, technical and operational expertise. There is relatively little precedent for this within the NHS, but in the US, early-adopter medical centres have independently developed organisational structures and multidisciplinary committees which enable and standardise these collaborative efforts.[24] Establishing these within NHS trusts is likely to be supportive of AI implementation efforts inside and outside of ophthalmology. It is also very common for these organisations to benefit from individuals who hold relatively superficial expertise across a breadth of relevant clinical, operational and technical domains. The need for such individuals has been called for in various UK policy documents since 2019.[41] So far in the NHS, it has largely been met by the provision of additional roles and training for a handful of clinicians, but more recent policy suggests that roles dedicated to this multi-disciplinary coordination function may be funded.[42-44] The continuation of this growing recognition of a distinct role and skillset within the NHS workforce is likely to support successful AI implementation and reduce NHS expenditure on third part consultancy.

With or without these AI-specialised staff or organisational structures, NHS trusts looking to adopt AI-enabled workflows will need to identify and bring together key stakeholders within their organisation. Without such collaborations it will be challenging for trusts to convene the perspectives and expertise necessary to account for the sociotechnical complexity of evaluating the suitability of available AIaMD for locally defined problems. These collaborations will include senior leadership representation, informatics and or information technology specialists, information governance specialists, legal specialists, relevant clinicians, procurement specialists and patient representatives.[24] This is critical to allow full characterisation of the problems that individual NHS trusts aim to address, and the nature of the evidence these local stakeholders require to trust and use AIaMD. This is likely to require evidence generation within the NHS trust environment and so internal or external funds to enable a tailored local evaluation and business case development will be important. The results of these evaluations must be made accessible to stakeholders to build trust and motivate adoption.[45] Connecting teams leading AI implementation across separate NHS trusts will also be valuable to enable the sharing of practical and credible insights between organisations to guide local expectations and investments.

8.6 Concluding remarks

The sight-threatening imbalance between demand and capacity in macula services is a persistent and broadly recognised problem for stakeholders in nAMD care. AI-enabled macula services are accepted by many as a potential solution and may deliver a variety of additional benefits. For now, it seems that the relative novelty of relevant AIaMD and the absence of evidence to inform and guide implementation strategies at the local level are the main factors that have prevented implementation. The outcomes of future implementation efforts will be strongly influenced by various stakeholder groups, many of whom have limited insight or contact with the others. Fortunately, considered design AIaMD, interventions and care pathways for nAMD appear able to mitigate, at least partially, against this complexity and disconnect between stakeholders. Such considered design seems able to offer a value proposition that satisfies all stakeholder groups.

The present thesis should not directly inform local efforts to implement AI-enabled macula services. However, it provides a blueprint for the design and conduct of local evaluations that can and justifies the investment they will require. It should also support policy makers as they consider which of their strategic priorities in ophthalmology are most amenable to AI innovation. It seems that the immediate next steps for implementation lie with early-adopter provider organisations such as NHS trusts. Their experiences, and the evidence that they generate and disseminate, will shape the future of AI-enabled macula services across the NHS.

8.7 References

1. Hogg, H.D.J., et al., Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence. J Med Internet Res, 2023. 25: p. e39742.

2. Al-Zubaidy, M., et al., Stakeholder Perspectives on Clinical Decision Support Tools to Inform Clinical Artificial Intelligence Implementation: Protocol for a Framework Synthesis for Qualitative Evidence. JMIR Res Protoc, 2022. 11(4): p. e33145.

3. Taribagil, P., et al., Integrating artificial intelligence into an ophthalmologist's workflow: obstacles and opportunities. Expert Review of Ophthalmology, 2023. 18(1): p. 45-56.

4. Camaradou, J.C.L. and H.D.J. Hogg, Commentary: Patient Perspectives on Artificial Intelligence; What have We Learned and How Should We Move Forward? Adv Ther, 2023. 40(6): p. 2563-2572.

5. Hogg, H., et al., Where should machine learning-enabled treatment monitoring be placedwithin the nAMD care pathway and what will influence its implementation?, in Annual Congress of the Royal COllege of Ophthalmologists. 2023: Birmingham.

6. Hogg, H.D.J., et al., Evaluating the translation of implementation science to clinical artificial intelligence: a bibliometric study of qualitative research. Front Health Serv, 2023. 3: p. 1161822.

7. Hogg. HDJ, et al., Safety and efficacy of an artificial intelligence-enabled decision tool for treatment decisions in neovascular age-related macular degeneration and an exploration of clinical pathway integration and implementation: protocol for a multi-methods validation study. BMJ Open, 2023. 13(2): p. e069443.

8. Hogg, H., Treatment decisions for known nAMDa non-inferiority study of AI vs consultant led care, in The Annual Congress of the Royal College of Ophthalmologists. 2023: Birmingham.

9. Hogg, H.D.J., et al., Unlocking the potential of qualitative research for the implementation of artificial intelligence-enabled healthcare. Journal of Medical Artificial Intelligence, 2023. 6.

10. De Fauw, J., et al., Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med, 2018. 24(9): p. 1342-1350.

11. Topol, E.J., High-performance medicine: the convergence of human and artificial intelligence. Nat Med, 2019. 25(1): p. 44-56.

12. Barclay, S., Smith, C., and Department of Health and Social Care, £21 million to roll out artificial intelligence across the NHS. 2023.

13. European Commission, EUDAMED - European Database on Medical Devices. 2023.

14. The Medicines and Healthcare products Regulatory Agency, Public Acces Registration Databse (PARD). 2023.

15. RetinAI, RetinAI and Retina Consultants of America Join Forces to build the Most Comprehensive U.S.-based Real World Evidence Database in Ophthalmology. 2023.

16. The Royal COllege of Ophthalmologists, Launch of physician associate pilot. 2023.

17. Kawamoto, K., et al., Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj, 2005. 330(7494): p. 765.

18. Damschroder, L.J., et al., The updated Consolidated Framework for Implementation Research based on user feedback. Implement Sci, 2022. 17(1): p. 75.

19. American National Standards Institute, Definitions/Characteristics Of Artificial Intelligence In Health Care - ANSI/CTA-2089.1-2020. 2020.

20. Ong, A., et al., Artificial Intelligence as a Medical Device for Ophthalmic Imaging in Europe, Australia, and the United States: Protocol for a Systematic Scoping Review of Regulated Devices. Open Science Framework, 2023.

21. Wilkinson, E., NICE approval of AI technology for radiotherapy contour planning. The Lancet Oncology, 2023. 24(9): p. e363.

22. The National Institute for health and Care Excellence, Artificial intelligence (AI) software to help clinical decision making in stroke. 2022.

23. Art, S., et al., IDEAL-D: a rational framework for evaluating and regulating the use of medical devices. BMJ, 2016. 353: p. i2372.

24. Kim, J.Y., et al., Organizational Governance of Emerging Technologies: AI Adoption in Healthcare, in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023, Association for Computing Machinery: Chicago, IL, USA. p. 1396–1417.

25. National Academy of Sciences, Reproducibility and Replicability in Science. 2019, Washington, DC: National Academies Press.

26. The National Institute for health and Care Excellence, Early value assessment interim statement; NICE process and methods [PMG39]. 2022.

27. Loewenstein, A., et al., Save our Sight (SOS): a collective call-to-action for enhanced retinal care across health systems in high income countries. Eye (Lond), 2023. 37(16): p. 3351-3359.

28. Ji Eun Diana, H., et al., Teleophthalmology-enabled and artificial intelligence-ready referral pathway for community optometry referrals of retinal disease (HERMES): a Cluster Randomised Superiority Trial with a linked Diagnostic Accuracy Study—HERMES study report 1—study protocol. BMJ Open, 2022. 12(2): p. e055845.

29. The National Institute for health and Care Excellence., AI technologies for detecting diabetic retinopathy. 2021.

30. The Software as a Medical Device Working Group, "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations 2014, International Medical Device Regulators Forum.

31. European Commission, Notified Bodies Survey on certifications and applications (MDR/IVDR). 2023.

32. Ross, A.H., et al., Recommendations by a UK expert panel on an aflibercept treat-and-extend pathway for the treatment of neovascular age-related macular degeneration. Eye, 2020. 34(10): p. 1825-1834.

33. Holz, F.G., et al., Does real-time artificial intelligence-based visual pathology enhancement of three-dimensional optical coherence tomography scans optimise treatment decision in patients with nAMD? Rationale and design of the RAZORBILL study. Br J Ophthalmol, 2023. 107(1): p. 96-101.

34. Chin-Yee, D., et al., A systematic review of as needed versus treat and extend ranibizumab or bevacizumab treatment regimens for neovascular age-related macular degeneration. Br J Ophthalmol, 2016. 100(7): p. 914-917.

35. Annastazia, E.L., et al., FENETRE study: quality-assured follow-up of quiescent neovascular age-related macular degeneration by non-medical practitioners: study protocol and statistical analysis plan for a randomised controlled trial. BMJ Open, 2021. 11(5): p. e049411.

36. Liu, X., et al., The medical algorithmic audit. The Lancet Digital Health, 2022. 4(5): p. e384-e397.

37. Vasey, B., et al., Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 2022. 28(5): p. 924-933.

38. Curran, G.M., et al., Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care, 2012. 50(3): p. 217-26.

39. Brown, C.H., et al., An Overview of Research and Evaluation Designs for Dissemination and Implementation. Annu Rev Public Health, 2017. 38: p. 1-22.

40. Sarkies, M.N., et al., Making implementation science more real. BMC Medical Research Methodology, 2022. 22(1): p. 178.

41. NHS England, The Topol Review. 2019.

42. NHS England. Topol Digital Fellowships. 2023; Available from: https://topol.hee.nhs.uk/digital-fellowships/.

43. Guy's and St Thomas' NHS Foundation Trust. Fellowship in Clinical Artificial Intelligence. 2023; Available from: https://gstt-csc.github.io/fellowship.html.

44. NHS AI Lab, NHS Transformation Directorate and NHS Health Education England, Developing healthcare workers' confidence in artificial intelligence (AI) (Part 2). 2022.

45. Kwong, J.C.C., et al., The silent trial - the bridge between bench-to-bedside clinical AI applications. Front Digit Health, 2022. 4: p. 929508.