Aspects of technology-based case-finding for eye disease

Bruno Miguel Rei Fidalgo

Submitted for the degree of

Doctor of Philosophy

City, University of London

Division of Optometry and Visual Science

January 2020

Contents

		Page Number
Contents		2
List of Tables		5
Lists of Figures		6
Acknowledgements		8
Declaration		9
Abstract		10
Abbreviations		12
Chapter 1: Introduction and objectives		
1.1	Introduction	15
1.2	Objectives	16
1.3	References	18
Chapter 2: Methodology and reporting of diagnostic accuracy studies of automated perimetry in glaucoma: evaluation using a standardised approach.		
2.1	Background	19
	• Introduction to diagnostic testing	19
	• Diagnostic accuracy studies	19
	• Methodological quality and bias in diagnostic accuracy studies	21
	• Quality of reporting of diagnostic accuracy studies	23
	• Diagnostic testing in opthalmology	27
2.2	Methods	28
	• Search Strategy	28
	• Study selection	28
	• Data extraction and management	28
	○ Assessment of methodological quality (QUADAS) and quality of reporting (STARD) of included studies	28
	○ Inter-rater reliability	29
	○ Evaluation of ophthalmic journal endorsement of reporting guidelines	29
	• Data synthesis and statistical analysis	29
2.3	Results	31
	• Results of search	31
	• Inter-rater reliability for QUADAS and STARD	31
	• Compliance with STARD 2003	33
	• Methodological Quality Assessment of Included Studies	38
	• Evaluation of ophthalmic journal endorsement of reporting guidelines	42

2.4 Discussion 44
- Strengths and limitations of this study 45

2.5 Role in the study 47
2.6 References 48
2.7 Published Article 52

Chapter 3: Development of a new glaucoma screening test using temporally modulated flicker

3.1 Introduction 61
3.2 Materials and Methods 64
- Development of the Accelerator 4 alternative forced choice flicker test prototype 64
  - Stimulus Design 64
  - Algorithm development 66
  - Algorithm performance simulation 67
  - Testing protocol 69
  - Selection of suitable threshold criteria 69
  - Assessment of individual reliability and learning effects 70

3.3 Results 71
- Determination of optimal threshold criteria 71
- Inherent Variability 73
- Assessment of individual reliability and learning effects 74

3.4 Commentary on the pre-clinical development of the A4FTp 76
3.5 Diagnostic accuracy of the A4FTp for the detection of COAG 79
- Purpose 79
- Methods 79
  - Reference standard 80
  - Inclusion and exclusion criteria 80
  - Index tests 80
  - Statistical Analysis 81
- Results 81
- Strengths and limitations of this study 86
- Conclusion 86

3.6 Role in the study 87
3.7 References 88
3.8 Published Article 92

Chapter 4: Diagnostic accuracy of structural and functional tests for the detection of sight-threatening eye diseases in a community setting

4.1 Background 104
4.2 Methods 106
- Recruitment of participants 106
- Description of index tests 107

- Frequency doubling technology (FDT) perimeter 107
- Moorfields motion displacement threshold test (MMDT) 108
- IVue Spectral domain OCT (SD-OCT) 110
- Ocular response analyser (ORA) 112

- Index test quality indicators and defined thresholds 114
- Reference standard exam 115
- Validation of the Reference standard examiner 115
- Index test acceptability questionnaire 118
- Sample size calculation 118
- Statistical analysis 119
  - Primary analysis: index test performance for the detection of COAG 119
  - Secondary analysis: index test performance for the detection of sight-threatening eye disease 119
- Analysis of user acceptability 120

4.3 Summary of main results 121

- Demographic and clinical characteristics of the study population 121
- Diagnostic accuracy of screening tests 123
  - Detection of COAG 123
  - Detection of any sight-threatening eye disease 125
  - Participant acceptability of screening tests 131

4.4 Discussion 133

- Strengths and limitations of the study 135

4.5 Role in the study 137
4.6 References 138
4.7 Published article 2015 141
4.8 Published article 2019 150

Chapter 5: Summary of results and directions for future work

5.1 Summary 158
5.2 Directions for future work 160
5.3 References 162

Appendix 1. Summary table of studies included in quantitative analysis 163
Appendix 2. 4AFC flicker algorithm 169
Appendix 3. Patient Acceptability Questionnaire 176
Appendix 4. Other Published Articles 179

List of Tables

		Page Number
Table 2.1	QUADAS 2003 items	22
Table 2.2	Sources of bias of diagnostic accuracy studies (adapted from Roever et al 2015)	23
Table 2.3	STARD 2003 Checklist	25
Table 2.4	OVID search strategy	30
Table 2.5	STARD item (scored as “Yes”)	34
Table 2.6	QUADAS items (scored as “Yes”)	39
Table 2.7	Results of audit of STARD and CONSORT inclusion in journal instructions to authors	43
Table 3.1	Inter-session modulation levels and test durations (per eye)	71
Table 3.2	Demographical and summary clinical data	82
Table 3.3	Pairwise comparison of ROC curves of the index tests	84
Table 4.1	OCT structural parameters	111
Table 4.2	A summary of index test quality indicators and cut-offs used in the study	114
Table 4.3	Diagnostic criteria used for classifying the most prevalent eye diseases	117
Table 4.4	Demographic data for the 505 participants	121
Table 4.5	Ocular pathology identified by the reference standard examination	122
Table 4.6	Sensitivity and specificity of index tests for detection of definite COAG	124
Table 4.7	Sensitivity at 90% specificity and Partial area under the receiver operation characteristic curve (AUROC) of index tests for detection of COAG	124
Table 4.8	Sensitivity and Specificity of individual screening tests for each Eye Disease	127
Table 4.9	Multivariate Stepwise Logistic Regression Analysis of Abnormal Screening Results with Ocular Disease	128
Table 4.10	Diagnostic performance of the optimised panel of screening tests (SD-OCT, FDT and measurement of visual acuity in identifying sight-threatening eye disease)	129

Lists of Figures

		Page Number
Figure 2.1	One-gate and Two-gate designs for a diagnostic accuracy study	20
Figure 2.2	Updated PRISMA flow diagram showing selection process	32
Figure 2.3	Stacked bar chart showing the STARD items reported chronologically by year of publication	36
Figure 2.4	STARD items 100% stacked bar before and after introduction of STARD in 2003	37
Figure 2.5	Bar chart showing the STARD items by Domain before and after 2003	38
Figure 2.6	Updated stacked bar chart showing the QUADAS items reported chronologically by year of publication	40
Figure 2.7	QUADAS items before and after introduction of STARD	41
Figure 2.8	Bar chart showing the percentage of QUADAS items scored as yes ordered by Domain before and after 2003	42
Figure 3.1	**A**. Glaucoma opportunistic screening pathway **B**. New glaucoma screening proposed pathway. Specialist evaluation	61
Figure 3.2	Flicker modulation stimulus within cosine envelope	63
Figure 3.3	**A**. Right-eye spatial arrangement for the 4AFC flicker test stimuli presented in a red field on a high refresh rate screen. **B**. Spatial location of the A4FTP flicker test locations for the right eye compared to the HFA 24-2 pattern. **C**. Field map by Wang and Henson (Wang et al, 2013).	65
Figure 3.4	Simulation of the performance of the Accelerator staircase method	67
Figure 3.5	Decilog modulation (Mod) levels ±1SD across the 4 sessions for each threshold criterion	72
Figure 3.6	Test duration (Dur) per eye in seconds ± 1SD across all 4 sessions for each threshold criterion	72
Figure 3.7	Decilog modulation (Mod) levels ±1SD over the four visual field locations, normalized to the group mean, for each threshold criterion	74
Figure 3.8	Trend analysis over 10 runs of the thresholds for each field location averaged for four control observers. **A**: Data for the criterion run length of	75

	8 (T8 condition). **B**: Data for the criterion run length of 12 (T12 condition)	
Figure 3.9	Study Flow Diagram	79
Figure 3.10	Histogram of the mean log thresholds of control and glaucoma subgroups. Glaucomatous visual field loss was classified using Hodapp-Parrish Anderson criteria (Hodapp et al 1993).	83
Figure 3.11	Sensitivity and specificity plot for A4FTp	83
Figure 3.12	Venn diagram of best performing parameter from the index tests in identifying the glaucoma cases alone or combined with the other tests	84
Figure 3.13	Likert responses from user acceptability survey	85
Figure 4.1	Study Flow Diagram. The number of eye conditions exceeds the number of participants due to ocular co-morbidity	107
Figure 4.2	Patel et al., 2000 scoring algorithm of FDT suprathreshold results	108
Figure 4.3	Moorfields motion displacement threshold test (MMDT) output	109
Figure 4.4	Vue SD-OCT scan outputs; iWellness (composite of GCC and retina map protocols) and ONH retinal nerve fibre layer thickness (RNFL)	112
Figure 4.5	Ocular Response Analyser (ORA) graphical plot	113
Figure 4.6	Index test acceptability questionnaire	118
Figure 4.7	Index test diagnostic effectiveness comparisons using receiver operating characteristic curves with sensitivity at set specificity estimates and associated 95% confidence intervals for detection of COAG	125
Figure 4.8	Venn diagrams presenting combined index test results for identification of A) COAG and B) other sight threatening eye diseases	130
Figure 4.9	Aggregated Likert scale responses to index test acceptability survey	131

Acknowledgements

I am deeply grateful to my supervisors Professor John Lawrenson for his never-ending patience, and Professor David Crabb for allowing me to start this project.

My gratitude to Priya Dabasia and Anish Jindal for their role in performing the reference standards and helping in the data analysis in chapter 3 and 4.

To my loving parents and brother for always being there for me, for their kind words and support this last decade.

Finally, to Theo, my partner, thank you for your support, love and putting up with me during my lowest times during this last few years.

Declaration

I grant powers of discretion to the Division of Optometry and Visual Science, City University London to allow this project to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to the normal consideration of acknowledgement.

Abstract

The aim of this thesis by prospective publication was to evaluate the comparative accuracy of screening technologies for the detection of COAG and other sight-threatening eye diseases.

It incorporates 4 peer-reviewed publications:

1). *Methodology and reporting of diagnostic accuracy studies of automated perimetry in glaucoma: evaluation using a standardised approach (Ophthalmic Physiol Opt. 2015; 35(3):315-23).* This publication reported on an evaluation of the methodological and reporting quality of diagnostic accuracy studies of perimetry in glaucoma and determined whether there had been any improvement since the publication of the Standards for Reporting of Diagnostic. Accuracy (STARD) guidelines in 2003. The main findings were that methodological and reporting quality was sub-optimal and appeared not to have improved substantially following the development of the STARD guidance.

2). *Development and validation of a new glaucoma screening test using temporally modulated flicker (Ophthalmic Physiol Opt. 2018; 38(6):617-628).* This publication described the psychometric characteristics and diagnostic accuracy of the Accelerator 4-Alternative Forced-Choice Flicker Test prototype (A4FTp) for detecting Chronic open angle glaucoma (COAG). The performance of the A4FTp was also compared with standard screening tests of ocular structure and function (Frequency Doubling Technology (FDT) perimetry and iVue Spectral Domain Optical Coherence Tomography (SD-OCT). The time taken to complete the A4FTp was relatively short and initial results are promising. The diagnostic accuracy for the A4FTp was comparable to those of the FDT and SD-OCT for detecting COAG and we concluded that with further refinement the A4FTp could have role in glaucoma detection in the future.

3). *Diagnostic accuracy of technologies for glaucoma case-finding in a community setting (Ophthalmology. 2015;122(12):2407-15).* This publication described a cross-sectional, observational, community-based study to evaluate the diagnostic performance of the FDT, Moorfields Motion Displacement Test (MMDT), iVue SD-OCT and ocular response analyzer (ORA) used alone or in combination, for the detection of COAG. Diagnostic performance of individual tests gave acceptable accuracy for COAG detection. The best performing parameter was inferior RNFL thickness recorded using the SD-OCT. Although the low specificity of visual-

function tests precluded their use in isolation, an acceptable performance was achieved by combining RNFL thickness analysis with visual function tests

4). *Role of advanced technology in the detection of sight-threatening eye disease in a UK community setting (BMJ Open Ophthalmology 2019; 4:e000347).* This publication extended the analysis described above to determine the performance of same screening tests for detecting sight-threatening eye disease in a cohort of elderly subjects recruited from primary care. The main finding was that a subset of screening tests (FDT, SD-OCT, together with a recorded visual acuity <6/12) was the most effective in detecting significant eye disease in this elderly population. The study provided useful preliminary data to inform the development of further larger, multi-center screening studies to validate this screening panel.

The work described in this thesis makes a useful contribution to the evidence base on the use of imaging and visual function technologies to identify COAG and other sight-threatening eye diseases in at-risk populations and provides clear directions for future research this area.

.

Abbreviations

AAC	Acute angle closure
A4FTp	Accelerator 4-Alternative Forced-Choice Flicker Test prototype
ACG	Angle closure glaucoma
AMD	Age-related macular degeneration
AREDS	Age-Related Eye Disease Study
AS-OCT	Anterior segment optical coherence tomography
ATD	Achromatic (A), red-green (Tritanope, "T") and blue-yellow (deuteranope, "D") Multichannel Functional Test
AUC	Area under the curve
AUROC	Area under the Receiver Operator Characteristic Curve
BCOVS	British Congress of Optometry and Vision Science
Cd m^{-2}	Candela per square meter
CH	Corneal hysteresis
CHRPE	Congenital hypertrophy of retinal pigment epithelium
CI	Confidence Interval
COAG	Chronic open angle glaucoma
CoO	College of Optometrists
CONSORT	Consolidated Standards of Reporting Trials
CoR	Coefficients of repeatability
CPU	Central Processing Unit
CRF	Corneal resistance factor
D	Diopter
DAP	Detection Acuity Perimetry
dB	Decibel
dL	Decilog
DOR	Diagnostic odds ratio
DMCO	Damato Multifixation Campimetry
DR	Diabetic retinopathy
ETDRS	Early Treatment Diabetic Retinopathy Study
EMP	Eye Movement Perimetry
EQUATOR	Enhancing the QUAlity and Transparency Of health Research
FDF	Flicker Defined Form Perimetry
FDI	Frequency-doubling illusion
FDT	Frequency Doubling Technology
FE	Fixation error
FLV	Focal loss volume
FN	False negative
FP	False positive
GAT	Goldmann Applanation Tonometer
GB	Gigabyte
GCC	Ganglion cell complex
GHT	Glaucoma Hemifield Test
GHz	Gigahertz

GLV	Global loss volume
GOC	General Optical Council
GP	General Practitioner
HFA	Humphrey Field Analyser
HRP	High-pass resolution perimetry
Hz	Hertz
IOP	Intraocular pressure
IQR	Interquartile range
IRMA	Intraretinal microvascular abnormality
ISGEO	International Society Geographical Epidemiological Ophthalmology
LED	Light-emitting diode
LN	Lower nasal
LOCS	Lens Opacity Classification System
LogMAR	Logarithm of the Minimum Angle of Resolution
LR	Late responses
LT	Lower temporal
MAP	Motion Automated Perimetry
MD	Mean deviation
MeSH	Medical Subject Heading
MMDT	Moorfields Motion Displacement Test
MP	Microperimetry
NA	Narrow angle
NCT	Non-contact tonometry
NICE	National Institute for Health and Care Excellence
NLR	Negative likelihood ratio
mmHg	Millimetres of mercury
NHS	National Health Service
NPV	Negative predictive value
NR	Not recorded
NSC	National Screening Committee
OCT	Optical Coherence Tomography
OHT	Ocular hypertension
OKP	Oculo-Kinetic Perimetry
ORA	Ocular Response Analyser
PAC	Primary angle closure
PACG	Primary angle closure glaucoma
PACS	Primary angle closure suspect
PERCEPT	PERformance CEntered Portable Test
PLR	Positive likelihood ratio
POAG	Primary open angle glaucoma
PP	Pulsar Perimetry
PPV	Positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PSD	Pattern standard deviation
PTD	Probability of true damage
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RAM	Random-Access Memory
RAP	Resolution Acuity Perimetry
RBP	Rarebit Perimetry
RCO	Royal College of Ophthalmologists
RCT	Randomised controlled trial
RGC	Retinal ganglion cells
RNFL	Retinal nerve fibre layer
RNIB	Royal National Institute of Blind People
ROC	Receiver Operator Characteristic Curve
s	Seconds
SAP	Standard Automated Perimetry
SD	Standard deviation
SD-OCT	Spectral-Domain Optical Coherence Tomography
SEM	Standard error of the mean
SITA	Swedish Interactive Thresholding Algorithm
SQI	Scan quality index
SROC	Summary receiver operator characteristic curve
SS-OCT	Swept-Source Optical Coherence Tomography
STARD	Standards for Reporting of Diagnostic Accuracy Studies
SWAP	Short Wavelength Automated Perimetry
TMP	Temporal Modulation Perimetry
TOP	Tendency-Oriented Perimetry
UCL	University College London
UK	United Kingdom
UN	Upper nasal
USA	United States of America
USB	Universal serial bus
UT	Upper temporal
VFE	Visual field extent
WCO	World Council of Optometry
WS	Waveform score
WHO	World Health Organization

Chapter 1.

Introduction and Objectives

1.1. Introduction

Visual impairment is an escalating global public health problem. Population growth coupled with an increasingly longevity in most countries, has led to rising numbers of people with visual impairment. In 2017, a systematic review and meta-analysis of available data sources, estimated that 216·6 million of the world's population was suffering from moderate or severe vision impairment (defined as visual acuity (VA) <6/18 but ≥3/60) with 36.0 million people blind (<3/60 in the better eye) (Flaxman et al 2017). The three leading causes of blindness were cataract, uncorrected refractive error, and glaucoma. In the case of moderate or severe vision impairment, main causes included; uncorrected refractive error, cataract, and age-related macular degeneration (AMD). Large regional variation in prevalence estimates were reported, with a higher prevalence of visual impairment due to AMD in high-income countries (Bunce et al 2015).

In the UK, visual impairment and blindness affects approximately 2 million people (3% of the population) (Pezzulu et al 2018) and the overall health and social care costs associated with sight loss and blindness have been estimated at £15.8 billion per annum (Pezzullu et 2018). Given that a significant proportion of sight loss and blindness is preventable, further investment in prevention, early detection and timely intervention is likely to be an effective strategy to reduce the burden of visual disability and improve socioeconomic outcomes (WHO 2013).

Glaucoma is a progressive optic neuropathy that can potentially lead to blindness if left untreated. In the UK, glaucoma remains the second most common cause of blindness and is responsible for 11% of cases of severe sight impairment (Quartilho et al 2016). Current models of detection of chronic open-angle glaucoma (COAG) rely on opportunistic case-finding and although there is evidence that in high-income countries at least 50% of cases remain undiagnosed (Klein et al 1992; Mitchell et al 1996), no country has so far introduced population screening for glaucoma. Health economic modelling studies conducted in Finland (Vaahtoranta-Lehtonen et al 2007) and the UK (Hernandez et al 2008) have suggested that screening for COAG could be cost-effective for specific subgroups at higher risk.

An ideal screening test for COAG should be quick, easy to perform and interpret, and be acceptable to the population being tested. It should also have sufficient

diagnostic power to distinguish between those who have and those who do not have COAG. There is currently a lack of high-quality diagnostic accuracy studies for COAG detection (Mowatt et al 2008; Michelessi et al 2015). Screening test accuracy is variable across studies and test performance is frequently overestimated due to the use of a case-control design that compares a healthy with a diseased population. Ideally, diagnostic accuracy studies should be carried out on patients selected consecutively at a defined stage of the clinical pathway.

The overall aim of the studies described in this thesis is to evaluate the comparative accuracy of screening technologies for the detection of COAG and other sight-threatening eye diseases. This includes an assessment of the methodological quality and adherence to reporting standards of existing diagnostic accuracy studies; the development and evaluation of a novel screening test for COAG that incorporates temporally modulated flicker and establishing the performance of a battery of conventional diagnostic tests for the detection of glaucoma and other sight-threatening eye diseases in a population of elderly subjects recruited from primary care.

1.2. Objectives

1. To assess the methodological quality of diagnostic test accuracy studies of using the QUADAS 2003 tool (an evidence-based quality assessment tool) and evaluate the accuracy and completeness of reporting of these studies based on the Standards for Reporting Diagnostic Accuracy (STARD) 2003 checklist (Chapter 2).
2. To develop a new algorithm to determine flicker sensitivity thresholds in susceptible areas of the visual field (Accelerator 4-Alternative Forced-Choice Flicker Test prototype (A4FTp)) that could be used as a rapid screening test for COAG (Chapter 3)
3. To evaluate the psychometric properties and diagnostic accuracy of the A4FTp for the detection of COAG (Chapter 3)
4. To assess the case-finding performance of structural (iVue Optical Coherence Tomography (OCT)) and functional (Frequency Doubling Technology Perimeter (FDT); Moorfields Motion Displacement Test (MMDT); Ocular Response Analyzer (ORA)) screening tests used alone or in combination for the detection of COAG in a cohort of elderly subjects recruited from primary care (Chapter 4)

5. To extend the analysis of the screening strategy outlined in 4., to determine the predictive value of an optimised panel of structural and functional tests to detect any sight-threatening eye disease (Chapter 4)

1.3. References

Bunce, C., Zekite, A., Walton, S., Rees, A., Patel, P.J., 2015. Certifications for sight impairment due to age related macular degeneration in England. Public Health 129, 138–142.

Flaxman, S. R. et al. (2017) 'Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis', The Lancet Global Health, 5(12), pp. e1221–e1234.

Hernández, R.A., Burr, J.M., Vale, L.D., 2008. Economic evaluation of screening for open-angle glaucoma. International Journal of Technology Assessment in Health Care 24, 203–211.

Klein, B.E.K., Klein, R., Sponsel, W.E., Franke, T., Cantor, L.B., Martone, J., Menage, M.J., 1992. Prevalence of Glaucoma. Ophthalmology 99, 1499–1504.

Michelessi, M., Lucenteforte, E., Oddone, F., Brazzelli, M., Parravano, M., Franchi, S., Ng, S.M., Virgili, G., 2015. Optic nerve head and fibre layer imaging for diagnosing glaucoma. Cochrane Database of Systematic Reviews. 11, CD008803.

Mitchell, P., Smith, W., Attebo, K., Healey, P.R., 1996. Prevalence of Open-angle Glaucoma in Australia. Ophthalmology 103, 1661–1669.

Mowatt, G., Burr, J.M., Cook, J.A., Siddiqui, M.A.R., Ramsay, C., Fraser, C., Azuara-Blanco, A., Deeks, J.J., 2008. Screening Tests for Detecting Open-Angle Glaucoma: Systematic Review and Meta-analysis. Investigative Opthalmology & Visual Science 49, 5373.

Pezzullo, L., Streatfeild, J., Simkiss, P., Shickle, D., 2018. The economic impact of sight loss and blindness in the UK adult population. Biomed Central Health Services Research 18.

Quartilho, A. et al. (2016) 'Leading causes of certifiable visual loss in England and Wales during the year ending 31 March 2013', Eye, 30(4), pp. 602–607.

Vaahtoranta-Lehtonen, H., Tuulonen, A., Aronen, P., Sintonen, H., Suoranta, L., Kovanen, N., Linna, M., Läärä, E., Malmivaara, A., 2007. Cost effectiveness and cost utility of an organized screening programme for glaucoma. Acta Ophthalmologica Scandinavica 85, 508–518

WHO (2013) Towards universal eye health: a global action plan 2014–2019 (WHA66.4) Available at: http://apps.who.int/gb/ebwha/pdf_files/WHA66-REC1/A66_REC1-en.pdf?ua=1 [Accessed 08/01/20]

Chapter 2.

Methodology and reporting of diagnostic accuracy studies of automated perimetry in glaucoma: evaluation using a standardised approach.

2.1. Background

Introduction to diagnostic testing

Diagnostic testing plays a key role in clinical decision-making, both in the identification of previously undiagnosed conditions and in the monitoring of existing conditions, including determining the response to therapy. Technological developments, particularly over the last two decades, have led to a significant increase in medical diagnostics across all specialties, largely driven by demands for earlier detection and increased speed and performance of the tests themselves (Nema et al 2014). In the effort to detect disease at its earliest clinical stage, diagnostic testing can be associated with unintended harms. These include false positives, leading to increased patient anxiety and unnecessary investigations. Testing can also result in over diagnosis and potentially unnecessary treatments (Holman et al 2017).

Diagnostic accuracy studies

Diagnostic test accuracy studies compare an 'index test' to a reference ('gold') standard, which is usually the best test available for accurately identifying the presence or absence of the condition of interest. There are two main types of diagnostic accuracy study (Figure 2.1): cross-sectional studies and diagnostic case control designs (also known as two-gate designs).

- In a cross-sectional study design all patients at risk of having the condition of interest, undergo the index and the reference test. At the time of inclusion in the study, there is clinical uncertainty about their disease status. Such studies are also referred to as 'single-gate' studies and are considered to more likely to provide a representative estimate of diagnostic test accuracy (Leeflang et al 2013).
- In a two-gate (case-control) design, subjects known to have the target condition are recruited and compared to healthy controls. Such studies are

prone to bias and may lead to an overestimation of test performance, particularly if only sections of the spectrum of disease and spectrum of non-diseased are included.

The performance of an index test is usually quantified by measures of diagnostic accuracy such as sensitivity and specificity, positive and negative predictive values, positive and negative likelihood ratios, the area under the ROC curve and diagnostic odds ratios (Šimundić 2009).

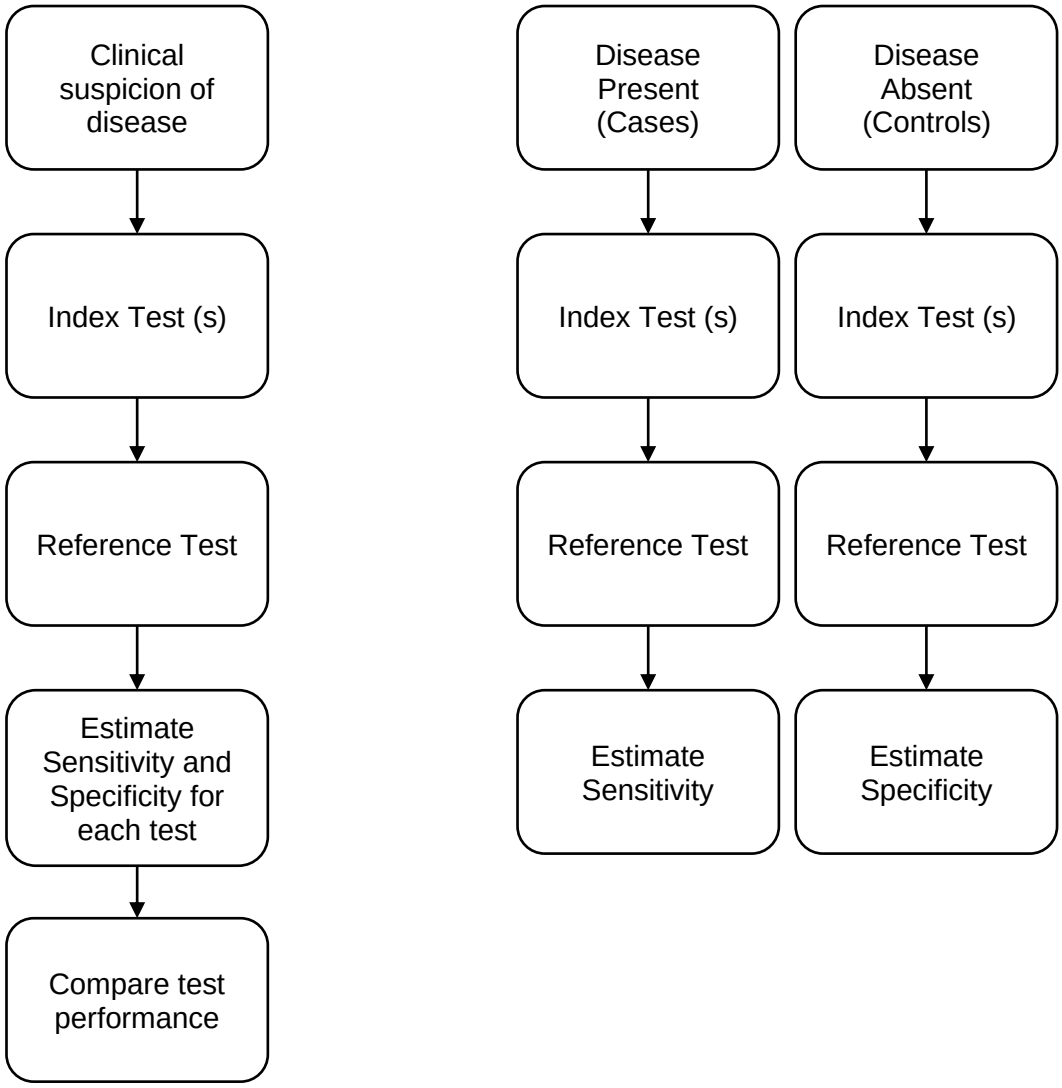


Figure 2.1. One-gate and Two-gate designs for a diagnostic accuracy study.

Methodological quality and bias in diagnostic accuracy studies

Diagnostic accuracy studies allow a clinician to make an informed decision regarding the potential utility of a new test. However, the evaluation of diagnostic test accuracy studies presents a number of challenges. The quality of a study is determined by its experimental design, the methods by which the study participants are recruited, the conduct of the index and reference tests and whether interpreters of the tests are masked. Overstating or understating results of new tests could lead to the premature adoption of a poorly performing test or delayed adoption of a high-quality test (Azuara-Blanco et al 2012).

The first Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool was created in 2003 (Table 2.1) and provides a useful checklist to assess the methodological quality (internal and external validity) of a diagnostic accuracy study. QUADAS is used by NICE and has been adopted by the Cochrane collaboration for the systematic reviews of diagnostic test accuracy. The tool is structured as a list of 14 items which are answered as ‘yes’, ‘no’ or ‘unclear’. The items cover the main sources of bias, including: spectrum/sampling bias, verification bias, disease progression bias, attrition bias further described in Table 2.2 as well as items asking about the execution of index and reference tests (Whiting et al 2003).

Table 2.1 – QUADAS 2003 items	
1	Was the spectrum of patient's representative of the patients who will receive the test in practice?
2	Were selection criteria clearly described?
3	Is the reference standard likely to correctly classify the target condition?
4	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?
5	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?
6	Did patients receive the same reference standard regardless of the index test result?
7	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?
8	Was the execution of the index test described in sufficient detail to permit replication of the test?
9	Was the execution of the reference standard described in sufficient detail to permit its replication?
10	Were the index test results interpreted without knowledge of the results of the reference standard?
11	Were the reference standard results interpreted without knowledge of the results of the index test?
12	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?
13	Were uninterpretable/ intermediate test results reported?
14	Were withdrawals from the study explained?

Table 2.2 – Sources of bias of diagnostic accuracy studies (adapted from Roever et al 2015)	
Source of bias	**Explanation**
Spectrum bias	When included patients do not represent the intended spectrum of severity for the target condition.
Classification bias	When the reference test does not correctly classify patients with the target condition.
Information bias	When the index results are interpreted with knowledge of the reference test results, or with more (or less) information than in practice
Verification bias	Ideally, all those who are tested with the index test should receive verification by the reference test (gold standard). Failure to do so can cause bias in accuracy estimates and is known as verification bias. Partial verification bias occurs when a proportion of patients do not undergo the reference test. Differential verification bias occurs when an alternative reference test is used to classify disease.
Attrition Bias	When withdrawals or uninterpretable test results are excessive or not adequately explained.
Disease progression bias	When the patients' condition changes between administering the index and reference test

For the original published study (Fidalgo et al 2015) and in the 2019 update described in this Chapter, we used the 2003 version of the QUADAS risk of bias tool, However, an improved version of the tool (QUADAS-2) has been developed (Whiting et al 2011) and is now the current standard for systematic reviews of diagnostic test accuracy. One of the major changes in QUADAS-2 is that the first three domains (patient selection, index test and reference test) as well as risk of bias, are also assessed in terms of concerns regarding applicability.

Quality of reporting of diagnostic accuracy studies

Reporting guidelines are tools developed to aid accurate, transparent and complete reporting of the key aspects of research studies, including a description of methods and findings. These guidelines are typically in the form of a checklist.

Probably the most commonly used and best-known reporting guideline is the CONSORT Statement (CONsolidated Standards Of Reporting Trials), which consists of a 25-item checklist of the essential items to include in a randomised controlled trial (RCT) report and a flow diagram template. The CONSORT statement was first published in 1996 and subsequently revised in 2001 and 2010 (CONSORT 2010). Since the development of this initial reporting guideline, there has been a proliferation of new reporting guidelines to improve the quality of published reports for most of the common types of study design.

The purpose of the STARD statement (Standards for Reporting of Diagnostic Accuracy Studies) was to develop a similar checklist to improve the completeness and transparency of reporting of studies of diagnostic accuracy (Bossuyt et al 2003). STARD consists of a list of 25 items covering the main sections of the paper and promoted the use of a flow diagram to present the study design and report the exact number of patients at each stage of the study (Table 2.3). The development of STARD started with an extensive search of several databases (MEDLINE, EMBASE, BIOSIS and methodological database from the Cochrane Collaboration). Following a review of relevant publications, a comprehensive list of items was created. A conference was organised with the aim of reducing the list of potential items, and to discuss the optimal format and phrasing of the checklist. A version of the checklist was field-tested and placed on a website with a request for comments. When all feedback was collected and considered, the STARD committee assembled the final version of the checklist and flow chart. STARD was disseminated via a series of publications in key journals together with guidance on the use of the checklist (Bossuyt et al 2003). The STARD statement was updated in 2015 (Bossuyt et al 2015). The update incorporated new information on sources of bias and made the checklist easier to use. The STARD 2015 statement increased the number of essential reporting items from 25 to 30. The same team have also developed a checklist for reporting essential items in journal or conference abstracts (Cohen et al 2017).

Table 2.3 - STARD 2003 Checklist	
	TITLE/ ABSTRACT
1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').
	INTRODUCTION
2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.
	METHODS
3	Describe the study population: the inclusion and exclusion criteria, setting and locations where the data were collected.
4	Describe participant recruitment: was recruitment based on presenting symptoms, results from previous tests or the fact that the participants had received the index tests or the reference standard?
5	Describe participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.
6	Describe data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?
7	Describe the reference standard and its rationale.
8	Describe technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.
9	Describe definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.
10	Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.
11	Describe whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.
12	Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals)
13	Describe methods for calculating test reproducibility, if done.

	RESULTS
14	Report when study was done, including beginning and ending dates of recruitment.
15	Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co-morbidity, current treatments, recruitment centres).
16	Report the number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).
17	Report time interval from the index tests to the reference standard, and any treatment administered between.
18	Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.
19	Report a cross-tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.
20	Report any adverse events from performing the index tests or the reference standard.
21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).
22	Report how indeterminate results, missing responses and outliers of the index tests were handled.
23	Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centres, if done.
24	Report estimates of test reproducibility, if done.
	DISCUSSION
25	Discuss the clinical applicability of the study findings.

Diagnostic testing in ophthalmology

The field of ophthalmology has seen significant developments in diagnostic testing. New tests are available that evaluate a variety of structural and functional ocular parameters. These allow clinicians to perform a more comprehensive eye examination, to assist in ruling in or ruling out a particular disease or providing prognostic information in those with established disease. Glaucoma presents a particular diagnostic dilemma. Epidemiological studies have shown that approximately half of glaucoma is undiagnosed in developed countries (Tielsch et al 1991; Klein et al 1992; Mitchell et al 1996; Quigley et al 1997; Weih et al 2001) and over 90% in developing countries (Ramakrishnan et al 2003; Vijaya et al 2008; Garudadri et all 2010; Budenz et al 2013). Whilst on the one hand it is essential to detect glaucoma before it causes visual disability (Johnson et al 2017), at the same time its low prevalence in the general population presents a challenge for case-finding. Even using tests with very high sensitivities and specificities will still result in a high proportion of false positives (Lawrenson et al 2014).

The aim of this study was to assess the methodological quality of diagnostic accuracy studies of perimetric tests in glaucoma, using the QUADAS tool and evaluate the accuracy and completeness of reporting of these studies based on the STARD checklist. We also investigated the possibility of an improvement in the quality of reporting since the publication of reporting guidelines.

2.2. Methods

Search Strategy

We used the OVID platform to search relevant bibliographic databases (MEDLINE, EMBASE, and Global Health) to identify diagnostic accuracy studies of perimetry published over a 25-year period between January 1993 and October 2019. The initial search (reported in Fidalgo et al 2015) was conducted in August 2013 and compared two time periods (1993-2003; 2004-2013), before and after publication of the original STARD checklist in 2003. An updated search was performed in October 2019, covering the period 2013-2019. The search terms used for both searches are described in Table 2.4. The search was limited to publications in the English language and studies performed on human subjects and included synonyms relating to perimetry, glaucoma and diagnostic test accuracy.

Study selection

For the published review, titles and abstracts from the bibliographic searches were screened by a single reviewer (BF) and for the 2019 update by two reviewers (BF and JL). Full-text articles were obtained for records judged relevant, or possibly relevant, by at least one review author. Reasons for exclusion were documented at this stage. Studies were included if they reported on measures of diagnostic accuracy of perimetry in glaucoma.

Data extraction and management

Assessment of methodological quality (QUADAS) and quality of reporting (STARD) of included studies

A data extraction form was prepared for each of the checklists in Microsoft Excel to score each item. A written justification for each judgement was included in the spreadsheet.

One reviewer (BF) assessed each of the included studies for methodological quality using the original 2003 QUADAS tool (Table 2.1) to assess the susceptibility to bias based on published guidance (Whiting 2003). All studies were evaluated, and the 14 questions were answered as ‘yes’, ‘no’ or ‘unclear’, corresponding to ’high’, ’low’ or ’unclear’ risk of bias (lack of information or uncertainty over the potential for bias) for each domain.

The same reviewer (BF) also assessed each of the included studies for quality of reporting using the 25 item STARD 2003 checklist (Table 2.3). The checklist includes items that are arranged under the following headings: (1) Title, abstract, and keywords (1 item), (2) Introduction (1 item), (3) Methods (11 items), (4) Results (11 items), and (5) Discussion (I item). Each item was graded by one reviewer (BF) as 'fully', 'partially', or 'not reported' according to predefined criteria (Bossuyt et al 2003) (Table 2.3). Items that were 'not applicable' were recorded as such. For example, given that perimetry is a non-invasive test the item 'Report any adverse events from performing the index tests or the reference standard' was excluded for all studies. We also recorded whether each study author cited the STARD checklist in the paper.

Inter-rater reliability

For the published review, two reviewers (JL and DC) independently rated QUADAS and STARD in a 20% random sample of included studies. An inter-rater reliability analysis was performed using the weighted Kappa statistic to determine consistency among reviewers (Landis & Koch 1977). For the update, BF rated all studies and these were then independently checked by JL. In all cases disagreements were resolved by consensus.

Evaluation of ophthalmic journal endorsement of reporting guidelines

In the originally published review, the 'Instructions for Authors' guidance of each of the Journals in which the articles were published was checked for references to STARD and whether adherence to STARD was a requirement for reporting diagnostic test accuracy studies. Following the updated search, this process was repeated to incorporate the new studies. For comparison, we also documented whether the instructions to authors referenced the CONSORT Statement, which describes the minimum set of recommendations for reporting randomized controlled trials that was first published in 1996.

Data synthesis and statistical analysis

Descriptive statistics were used to present the number and proportion of studies scoring 'yes' for each QUADAS domain for the periods before and after publication of the STARD tool in 2003. The overall adherence to STARD was calculated as the mean number and percentage of reported items for each item for the periods 1993-2003 and 2004-2019. An independent samples t-test was used to analyse differences in the percentage of reported STARD and QUADAS items for periods before and after publication of the guidelines.

Table 2.4. OVID search strategy	
1.	Perimetry
2.	Perimeter
3.	Standard Automated Perimetry
4.	SAP
5.	Visual field
6.	Motion displacement test
7.	Frequency doubling technology
8.	Flicker Defined Form
9.	High pass resolution
10.	HRP
11.	OKP
12.	Humphrey
13.	Henson
14.	Octopus
15.	Heidelberg
16.	Dicon
17.	Medmont
18.	Rarebit
19.	Ophthimus
20.	or/1-19
21.	Glaucoma$
22.	20 and 21
23.	Diagnostic Accuracy
24.	Diagnostic performance
25.	Precision
26.	ROC
27.	Receiver operating characteristic
28.	Sensitivity and specificity
29.	Sensitivity
30.	Specificity
31.	Diagnostic odds ratio
32.	DOR
33.	Area under the receiver operating characteristic curve
34.	AUC
35.	Likelihood ratio.
36.	or/23-35
37.	22 and 36
38.	Early diagnosis
39.	Differentiate
40.	Identify
41.	Detect
42.	Diagnosis
43.	Screening
44.	Case finding
45.	or/38-44
46.	22 and 45

2.3. Results

Results of search

The initial search, conducted in August 2013, identified 488 articles of which, 58 were eligible for inclusion in the published review (Fidalgo et al 2015). The search was updated in October 2019. This yielded 782 articles, 647 articles were excluded during title and abstract screening, as they were either duplicates (N=4), not written in English (N= 1) or failed to meet the inclusion criteria due to: not glaucoma (N= 19), not perimetry (N=403), did not report a diagnostic accuracy study (N=92), were reviews, conference abstracts, letters or notes or editorials or surveys (N=128). One hundred and thirty one articles were selected for full text screening. Following the assessment of the full text of the articles, a further 55 were excluded. 76 studies were included in the updated analysis covering the period 1993-2019. Figure 2.2 shows the PRISMA flow diagram. Details of the included studies in terms of the index and reference test are provided in Appendix 1. The most commonly used reference standard was a combination of optic disc examination, intraocular pressure measurement and standard automated perimetry (used in 21% of studies) an additional 16% also included gonioscopy.

Inter-rater reliability for QUADAS and STARD

Assessment of inter-rater reliability using the weighted kappa statistic, showed substantial agreement between reviewers. Kappa values for QUADAS and STARD were 0.70 and 0.81 respectively.

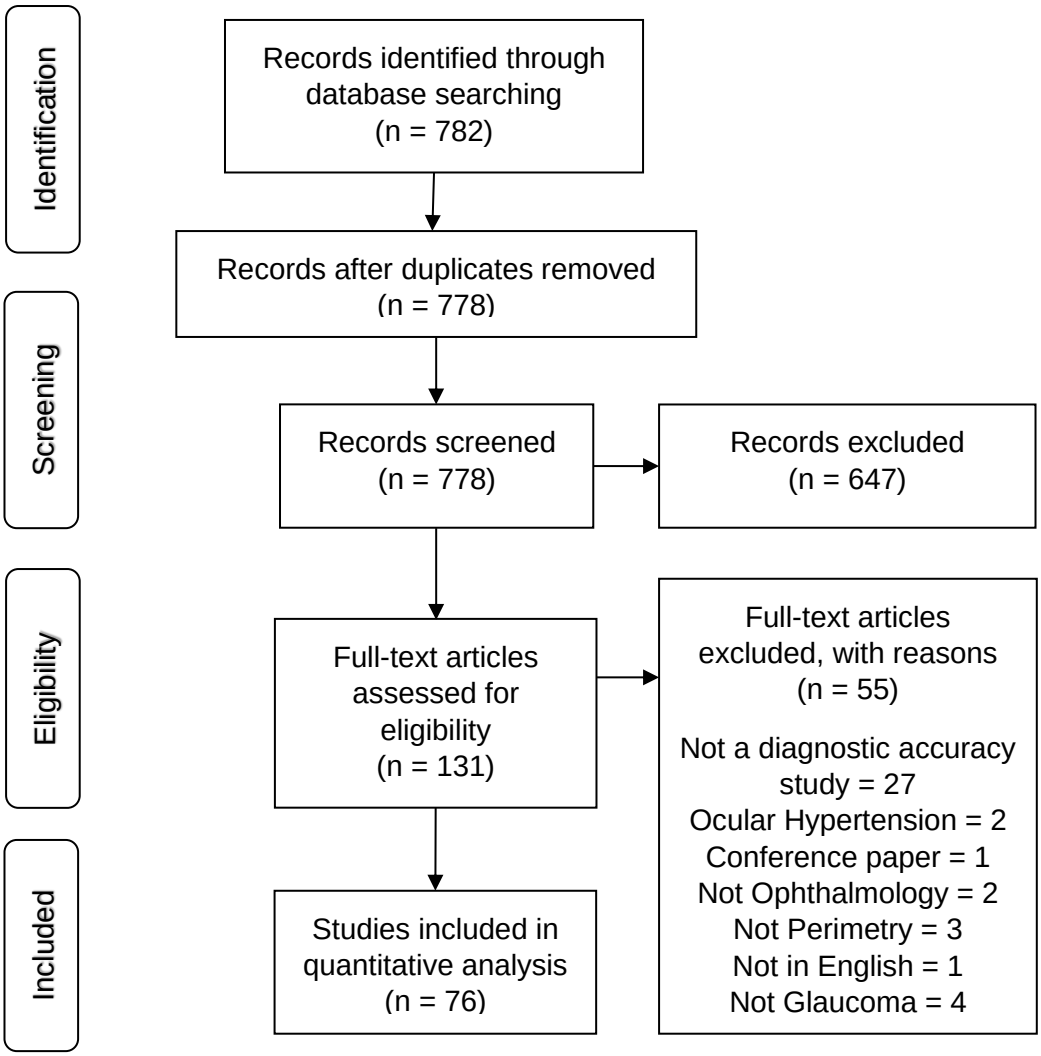


Figure 2.2 - Updated PRISMA flow diagram showing selection process

Compliance with STARD 2003

Given that perimetry is non-invasive, the STARD item 'Adverse effects from performing index or reference tests' was judged to be not applicable. The overall adherence to STARD based on 24 items was 51.5% (IQR 27.5 to 73.5%). Table 2.5 presents the adherence to STARD for each item for studies published between 1993 and 2019. Significant variability in the quality of reporting was observed, with the percentage of items being reported ranging from 7-100%. The most consistently reported items included: stating the research questions/study aims (item 2), description of participant recruitment (item 4), and discussion of the clinical applicability of the study findings (item 25). The most inconsistently reported items included: a description of whether the readers of the index tests and reference standard were masked to the results of the other test (item 11), a description of methods for test reproducibility/variability (item 13) and reporting of estimates of test reproducibility (item 24).

Figure 2.3 shows the STARD assessment of each article arranged chronologically by year of publication. The median score for articles published before and after the development of STARD was similar: 1993-2003, median=11 (range 6 to 15); 2003-2019 median=13 (range 8 to 22). The percentage of reported items showed no overall improvement between the two reporting periods (Figure 2.4) (P = 0.1693). Although items relating to the 'Introduction' and 'Discussion' were consistently reported pre and post STARD, (Figures 2.4 and 2.5). The publication of the STARD reporting guidelines appear not to have substantially improved in the reporting of items pertinent to the methods and results sections (Figure 2.5). Only three of the 76 studies assessed explicitly mentioned the use of STARD for preparing the manuscript.

Table 2.5. STARD item (scored as “Yes”)					
Item		**1993 to 2003 N (%)**	**2004 to 2019 N (%)**	**Total %**	**Difference between time periods %**
		N = 22	**N = 54**	**N = 76**	
TITLE/ ABSTRACT					
1*	Identify the article as a study of diagnostic accuracy	18 (82)	49 (90)	87	8
INTRODUCTION					
2	State the research questions of study aims.	22 (100)	54 (100)	100	0
METHODS					
3	The study population	4 (18)	22 (42)	34	24
4	Participant recruitment	22 (100)	54 (100)	100	0
5	Participant sampling	16 (73)	36 (63)	66	-10
6	Data collection strategy	8 (36)	32 (54)	49	18
7	The reference standard	6 (27)	37 (69)	56	42
8	Technical specifications of material and methods	13 (59)	46 (83)	76	24
9	Definition of and rationale of the index tests and the reference standard	10 (45)	41 (73)	64	28
10	The number, training and expertise of the persons executing and reading the tests	7 (32)	12 (21)	24	-11
11	Whether or not the tests were masked to the readers	2 (9)	11 (21)	17	12
12$	Methods for calculating and comparing measures of diagnostic accuracy and uncertainty	2 (9)	13 (23)	19	14
13	Methods to calculate reproducibility	1 (5)	5 (18)	7	13

RESULTS					
14	When the study was done	3 (14)	22 (42)	33	28
15	Clinical and demographic characteristics	13 (59)	37 (77)	71	18
16	Participant flow	5 (23)	21 (35)	31	12
17	Time interval between index and reference standard	3 (14)	14 (23)	20	9
18	Distribution of severity of disease and other diagnoses	18 (82)	44 (88)	86	6
19	Cross tabulation of the results of the index tests	12 (55)	36 (65)	61	10
20	Adverse effects from performing index or reference tests	NA	NA	NA	NA
21$	Estimates of diagnostic accuracy and measures of uncertainty	3 (14)	24 (44)	34	30
22	How indeterminate results, missing responses and outliers of the index tests were handled	6 (27)	20 (38)	34	11
23**	Estimates of diagnostic accuracy between subgroups	11 (50)	29 (58)	56	8
24	Estimates of test reproducibility	2 (9)	6 (10)	10	1
DISCUSSION					
25	Discuss the clinical applicability of the study findings	22 (100)	54 (100)	100	0

NA - not applicable, due to the non-invasive nature of the test.

* Considered to be positive if the words diagnostic accuracy appeared in the title or abstract, or if the article was identified using the MeSH term sensitivity and specificity.

** Considered to be NA if there were no subgroups.

$ If only estimates of diagnostic accuracy without a measure of uncertainty were given, this was scored as “partial.”

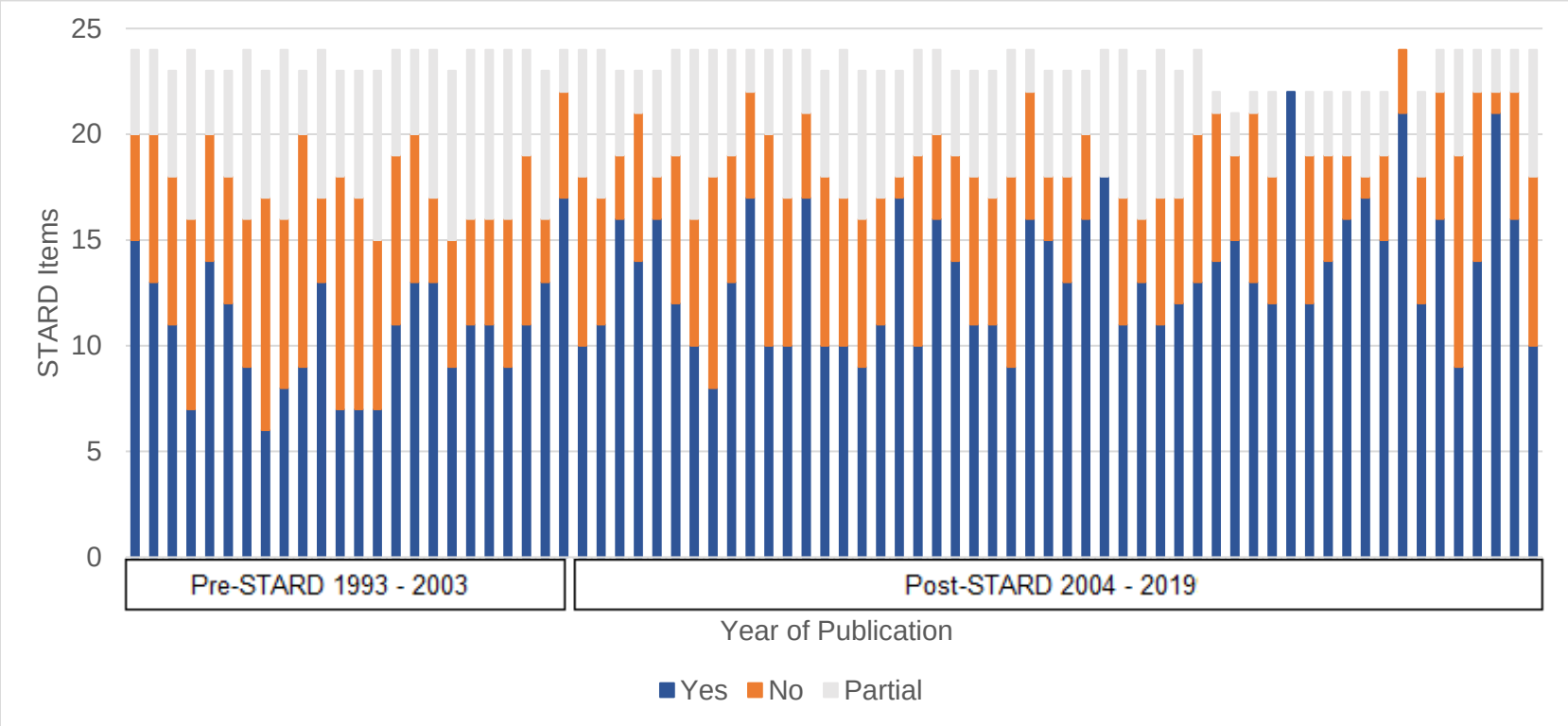


Figure 2.3. Stacked bar chart showing the STARD items reported chronologically by year of publication

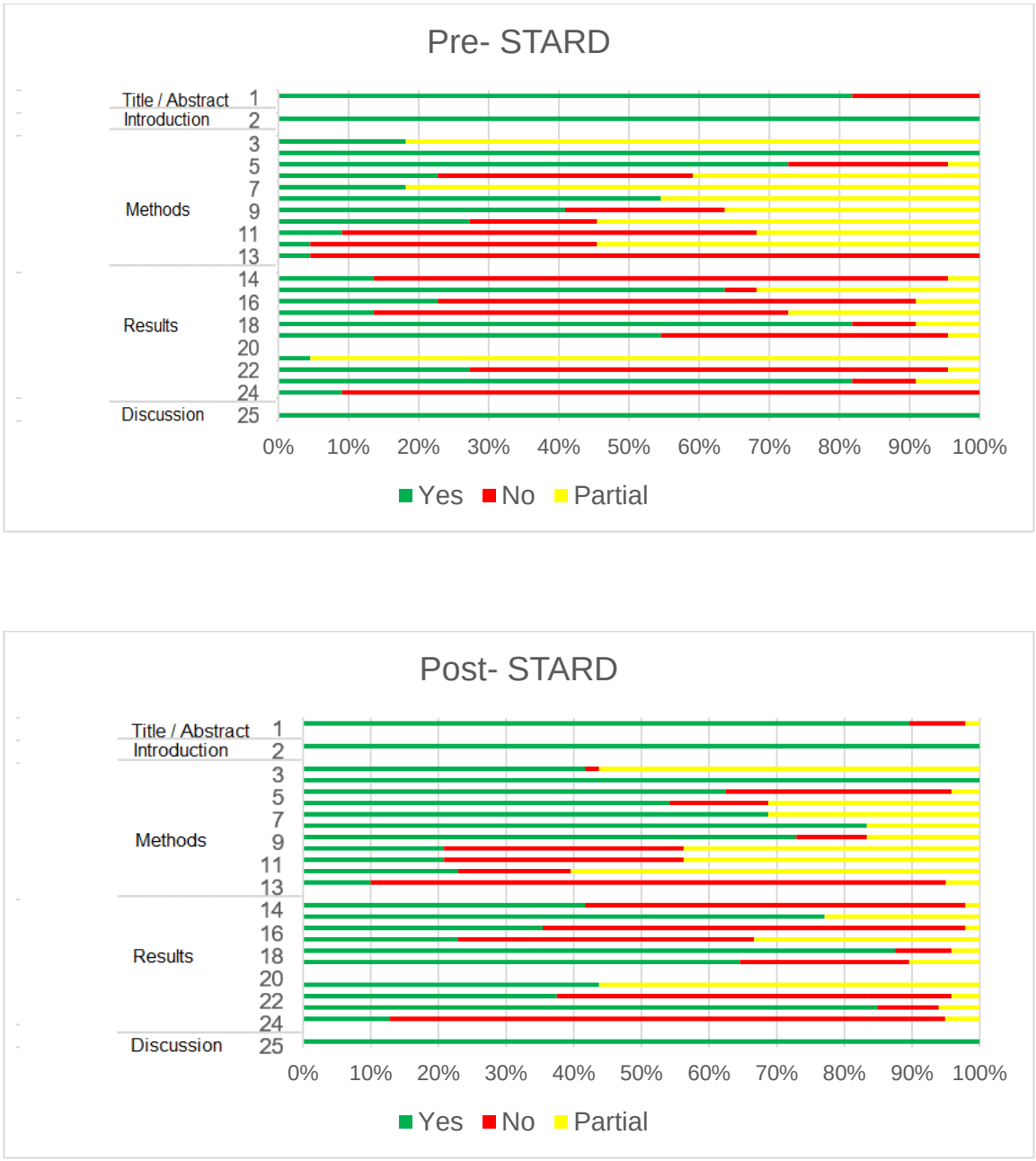


Figure 2.4 – STARD items 100% stacked bar before and after introduction of STARD in 2003

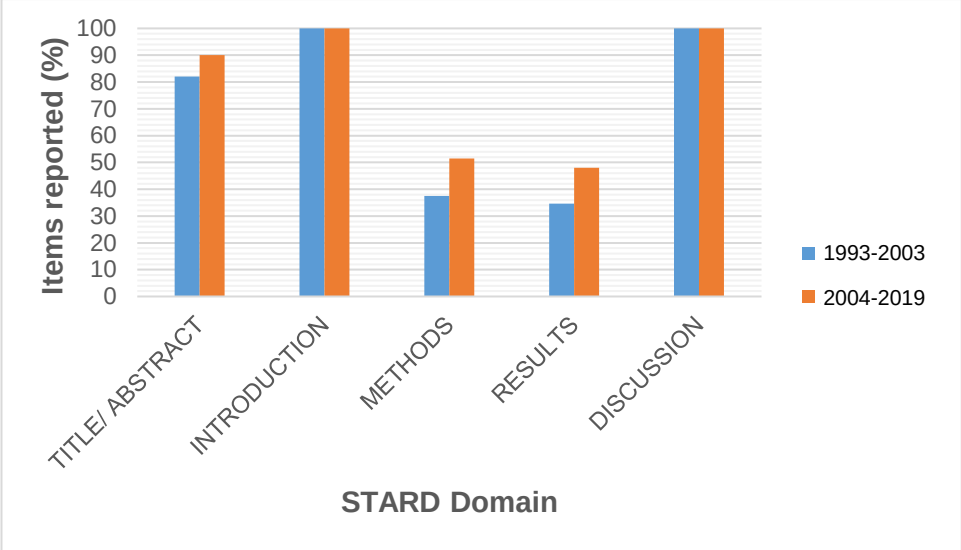


Figure 2.5. Bar chart showing the STARD items by Domain before and after 2003

Methodological Quality Assessment of Included Studies

The overall median number of QUADAS items scored as 'Yes' for articles published between 1993 and 2019 (N=76), was 9 (range 2 to 14). The median score for articles published before and after the development of STARD was similar: 1993-2003, median=8 (range 2 to 13); 2003-2019 median= 9 (range 4 to 14). Table 2.6 summarises the results for each item for each time-period and Figure 2.6 shows the QUADAS assessment of each article arranged chronologically by year of publication and in Figure 2.7 grouped items into risk of bias domains. The percentage of reported items showed no overall improvement between the two reporting periods (P = 0.1234), although specific items relating to the reference test and the flow and timing of the study showed a modest improvement (Figure 2.7).

Table 2.6. QUADAS items (scored as “Yes”)					
	Item	**1993 to 2003 N (%)**	**2004 to 2019 N (%)**	**Total (%)**	**Difference between time periods %**
		N = 22	**N = 54**	**N = 76**	
1	Was the spectrum of patients representative of the patients who will receive the test in practice?	9 (41)	19 (38)	43	-3
2	Were selection criteria clearly described?	18 (82)	51 (94)	90	12
3	Is the reference standard likely to correctly classify the target condition?	17 (77)	43 (81)	79	4
4	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	2 (9)	24 (50)	40	41
5	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	12 (55)	50 (92)	76	37
6	Did patients receive the same reference standard regardless of the index test result?	13 (59)	43 (77)	66	18
7	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	21 (95)	50 (92)	93	-3
8	Was the execution of the index test described in sufficient detail to permit replication of the test?	15 (68)	43 (77)	71	9
9	Was the execution of the reference standard described in sufficient detail to permit its replication?	7 (32)	40 (71)	53	39
10	Were the index test results interpreted without knowledge of the results of the reference standard?	4 (18)	17 (31)	28	13
11	Were the reference standard results interpreted without knowledge of the results of the index test?	8 (36)	24 (46)	47	10
12	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	17 (77)	39 (77)	79	0
13	Were uninterpretable/ intermediate test results reported?	6 (27)	29 (56)	48	
14	Were withdrawals from the study explained?	7 (32)	18 (33)	31	1

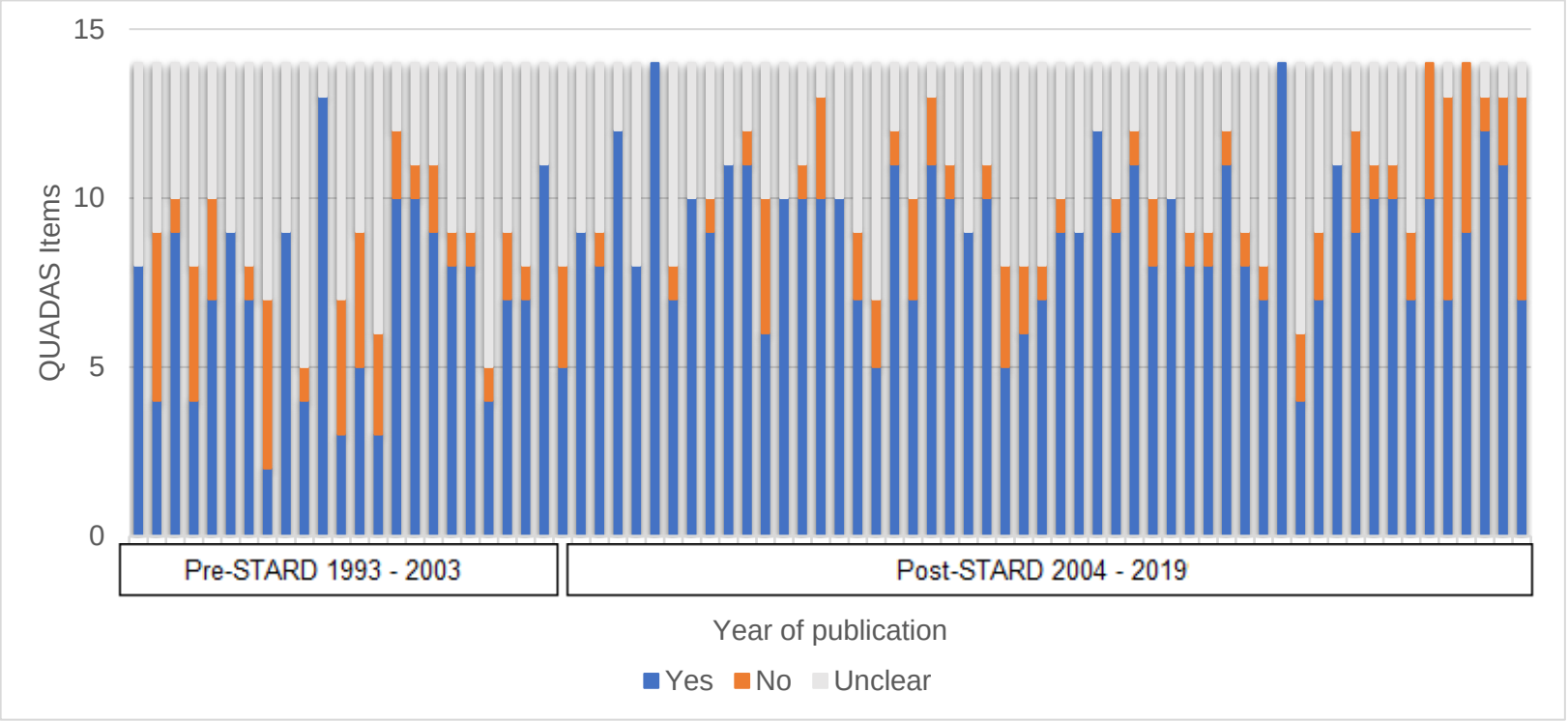


Figure 2.6. Updated stacked bar chart showing the QUADAS items reported chronologically by year of publication

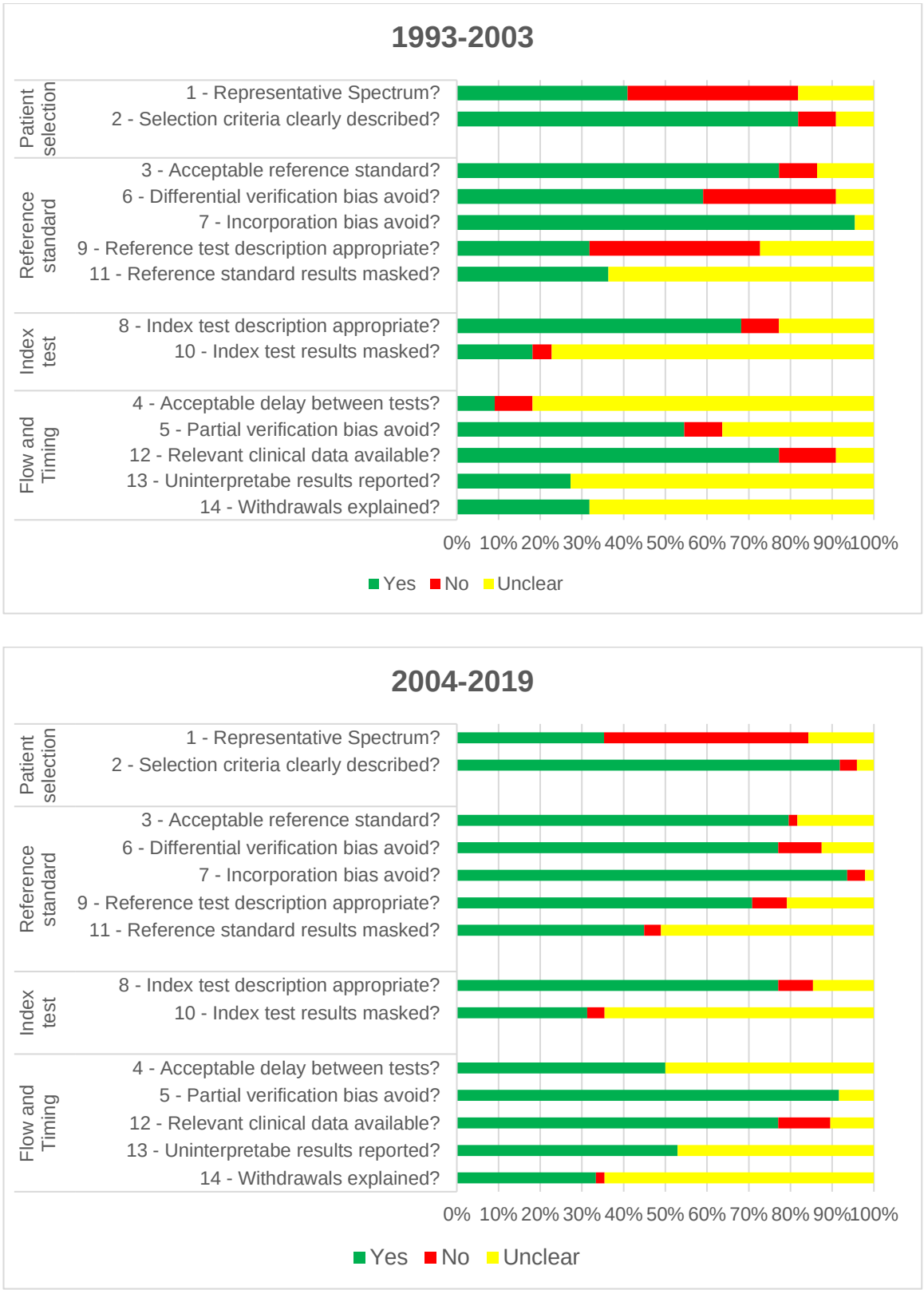


Figure 2.7 – QUADAS items before and after introduction of STARD

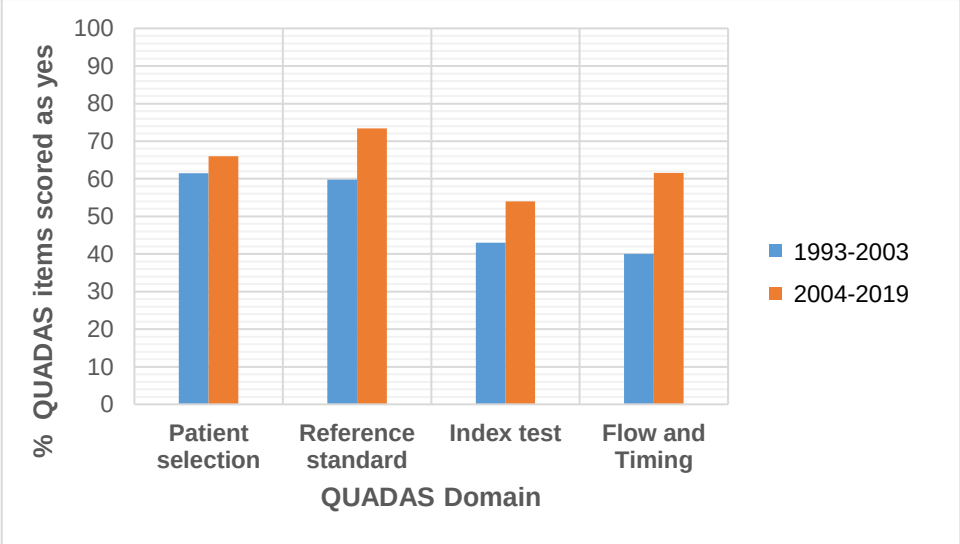


Figure 2.8. Bar chart showing the percentage of QUADAS items scored as yes ordered by Domain before and after 2003

Evaluation of journal endorsement of reporting guidelines

The included papers were published in 18 different journals. We were unable to access instructions for authors for two of the included journals: Bulletin de la Societe Belge d'Ophthalmologie and the African Journal of Medicine and Medical Sciences. Only 3 (19%) of these journals recommended that authors should use the STARD checklist and flow diagram for reports of diagnostic test accuracy. By comparison, 56% of journals stipulated that authors should use CONSORT when submitting a paper reporting a randomised controlled trial (Table 2.7).

Table 2.7. Results of audit of STARD and CONSORT inclusion in journal instructions to authors.

JOURNAL	**STARD**	**CONSORT**
Eye	N	Y
Investigative Ophthalmology & Visual Science	N	N
Acta Opthalmologica	N	N
British Journal of Ophthalmology	Y	Y
Japanese Journal of Ophthalmology	N	N
American Journal of Ophthalmology	N	Y
Journal of Glaucoma	N	N
Archives of Ophthalmology (now JAMA Ophthalmol)	Y	Y
Singapore Medical Journal	N	N
Ophthalmology	N	Y
Graefe's Archive for Clinical and Experimental Ophthalmology	N	N
European Journal of Ophthalmology	Y	Y
Ophthalmic Epidemiology	N	Y
Optometry and Vision Science	N	Y
Clinical and Experimental Ophthalmology	N	N
Canadian Journal of Ophthalmology	N	Y

2.4. Discussion

The aim of this study was to assess the quality of reporting and overall methodological quality of diagnostic accuracy studies that used perimetry to detect functional vision loss in glaucoma. We further investigated the impact of the publication of STARD on the quality of reporting by comparing articles published before and after the development and dissemination of the STARD checklist in 2003. The initial search and analysis was carried out in 2013, covering the period 1993-2013. This allowed a comparison of published studies over two 10-year time periods pre- and post STARD. We used the QUADAS tool to assess methodological quality and hypothesised that any improvements in reporting of diagnostic test accuracy studies would potentially lead to a corresponding improvement in overall methodological quality. The results of the initial analysis (1993-2013) were published in 2015 (Fidalgo et al 2015). In the same year, a revised 30-item STARD checklist was published (Bossuyt et al 2015). We subsequently updated the review in October 2019 to include studies that had been published since 2013. This identified a further 18 studies.

Across the whole of the evaluation period, the overall compliance with STARD was poor, with only 50% of the items adequately reported. A comparison of the studies published before and after the development of the of the original STARD checklist in 2003 found that overall, reporting had not substantially improved, with no further significant improvement since the publication of the updated checklist in 2015. Significantly, only three of the included studies reported that they had used the STARD checklist in the development of their manuscript. This is possibly not surprising since less than 20% of the included journals recommended in their submission instructions that authors should use STARD when reporting diagnostic test accuracy studies. By contrast, approximately 50% of the same journals referenced CONSORT for reporting RCT's. Our results, showing poor adherence to the original STARD checklist were consistent with previous studies in ophthalmology (Siddiqui et al 2005; Shunmugem et al 2006; Johnson et al 2007; Paranjothy et al 2007; Zafar et al 2008; Castillo et al 2014) and in other medical specialities (Coppus et al 2006; Wilczynsky et al 2008).

Given the poor reporting of the included studies, it is possibly not surprising that there was no statistical overall improvement in methodological quality (as judged by QUADAS on published studies before and after 2003). Intuitively, we would expect that even a well-conducted study would score poorly on a quality

assessment tool if the methods and results were not reported in sufficient detail. However, a recent study (Michelessi et al 2017) examined the relationship between STARD and QUADAS in a large set of studies on glaucoma. The authors reported that the relationship between the two tools was partial and difficult to interpret. Furthermore, they suggested that raters were using substantial context-specific knowledge when conducting a QUADAS assessment.

In the updated analysis, we examined QUADAS items grouped into four risk of bias domains. This analysis revealed a small to moderate improvement in some domains with the largest improvement in the timing and flow domain. This domain includes items relating to withdrawals and uninterpretable test results, evaluating partial verification bias and inappropriate delays between performing index and reference tests.

Over the past 20 years, reporting guidelines have been developed for a variety of study designs to assist health researchers in writing up their work for publication. These guidelines specifying a minimum set of items that are required to provide a transparent account of what was done together with a clear description of the results of the study. The Enhancing the Quality and Transparency of Reporting (EQUATOR) Network is an international organisation that aims to improve the quality of reporting of health research (Equator Network 2018). One of the goals of the organisation is to promote the use of reporting guidelines and support journals in implementing them. The Network maintains an online library of reporting guidelines for the main study types. There is an accumulating body of evidence to suggest that Journal endorsement of CONSORT improves the reporting of RCTs, although reporting is still sub-optimal (Turner et al 2012). Although it is unclear whether the poor adoption of STARD by ophthalmology journals is the primary reason for the incomplete reporting of diagnostic accuracy studies identified in this and other studies in ophthalmology. However, the experience from CONSORT would suggest that the implementation of STARD and enforcement by journal editors is likely to lead to quality improvements in the reporting of these studies.

Strengths and limitations of this study

This study used systematic methods to evaluate methodological quality and standards of reporting of diagnostic accuracy studies that were published in a defined area of ophthalmology. Our findings have added to the body of literature highlighting poor methodological quality and poor reporting of diagnostic accuracy studies across many health specialities. Despite the availability of STARD

reporting standards, there is clear evidence that these are not being followed. It is possible that the same advocacy strategies that were successful in increasing the adoption of the CONSORT guidelines for the reporting of RCTs could be applied to STARD. Much of the responsibility lies with journal editors, who should insist that STARD reporting guidelines are followed by requiring a completed checklist to be included with each paper submission.

We acknowledge a number of methodological limitations of the current study including:

1. Only assessing articles published in English
2. Using a single reviewer to screen all titles and abstracts and 80% of the QUADAS and STARD ratings.

Since the starting work on the review, updated STARD guidance has been produced (Bossuyt et al 2015). However, given that we were comparing studies published following the development of STARD in 2003, this early version was used throughout. Similarly, QUADAS has also been updated (Whiting 2011) and this version has now been adopted by organisations such as NICE and the Cochrane collaboration. However, at the point that we began the study, QUADAS 2 had not been universally adopted.

2.5. Role in the study

Under the supervision of Professors Lawrenson and Crabb, I wrote the protocol for the systematic review, conducted the bibliographic searches, screened the titles and abstracts for potentially included studies, extracted data and conducted the STARD and QUADAS assessments. I also wrote the first draft of the manuscript that was published in Ophthalmic and Physiological Optics. For the updated review described in this Chapter, I repeated the searches, extracted and analysed the data from the newly included studies.

My contribution: 100%

2.6. References

Abreu-Gonzalez, R., Gonzalez-Hernandez, M., Pena-Betancor, C., Rodriguez-Esteve, P., Gonzalez De La Rosa, M., 2018. New visual field indices of disharmony for early diagnosis of glaucoma, alone or associated with conventional parameters. European Journal of Ophthalmology 28, 590–597.

Azuara-Blanco, A., 2012. Diagnostic technologies in ophthalmology. Bentham Books, Saif Zone, Sharjah, U.A.E.

Baker, H., Ratnarajan, G., Harper, R.A., Edgar, D.F., Lawrenson, J.G., 2016. Effectiveness of UK optometric enhanced eye care services: a realist review of the literature. Ophthalmic and Physiological Optics 36, 545–557.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Moher, D., Rennie, D., de Vet, H.C.W., Lijmer, J.G., Standards for Reporting of Diagnostic Accuracy, 2003. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Annals of Internal Medicine 138, W1-12.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C.W., Kressel, H.Y., Rifai, N., Golub, R.M., Altman, D.G., Hooft, L., Korevaar, D.A., Cohen, J.F., 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. British Medical Journal h5527.

Budenz, D.L., Barton, K., Whiteside-de Vos, J., Schiffman, J., Bandi, J., Nolan, W., Herndon, L., Kim, H., Hay-Smith, G., Tielsch, J.M., Tema Eye Survey Study Group, 2013. Prevalence of glaucoma in an urban West African population: the Tema Eye Survey. JAMA Ophthalmology 131, 651–658.

Castillo, M.M., Mowatt, G., Lois, N., Elders, A., Fraser, C., Amoaku, W., Burr, J.M., Lotery, A.J., Ramsay, C.R., Azuara-Blanco, A., 2014. Optical coherence tomography for the diagnosis of neovascular age-related macular degeneration: a systematic review. Eye (London) 28, 1399–1406.

Cohen, J.F., Korevaar, D.A., Gatsonis, C.A., Glasziou, P.P., Hooft, L., Moher, D., Reitsma, J.B., de Vet, H.C., Bossuyt, P.M., STARD Group, 2017. STARD for Abstracts: essential items for reporting diagnostic accuracy studies in journal or conference abstracts. British Medical Journal 358, j3751.

CONSORT Transparent Reporting of Trials. (2010), http://www.consort-statement.org/, accessed 8/01/20.

Coppus, S.F.P.J., van der Veen, F., Bossuyt, P.M.M., Mol, B.W.J., 2006. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. Fertility and Sterility. 86, 1321–1329.

Fidalgo, B.R., Jindal, A., Tyler, C.W., Ctori, I., Lawrenson, J.G., 2018. Development and validation of a new glaucoma screening test using temporally modulated flicker. Ophthalmic and Physiological Optics 38, 617–628.

Garudadri, C., Senthil, S., Khanna, R.C., Sannapaneni, K., Rao, H.B.L., 2010. Prevalence and risk factors for primary glaucomas in adult urban and rural populations in the Andhra Pradesh Eye Disease Study. Ophthalmology 117, 1352–1359.

Hofmann, B., Welch, H.G., 2017. New diagnostic tests: more harm than good. British Medical Journal 358, j3314.

Johnson, C.A., Thapa, S., George Kong, Y.X., Robin, A.L., 2017. Performance of an iPad Application to Detect Moderate and Advanced Visual Field Loss in Nepal. American Journal of Ophthalmology. 182, 147–154.

Johnson, Z.K., Siddiqui, M.A.R., Azuara-Blanco, A., 2007. The quality of reporting of diagnostic accuracy studies of optical coherence tomography in glaucoma. Ophthalmology 114, 1607–1612.

Jones, S., Schultz, T., Enuameh, Y.A.K., Ovid Technologies, I., 2011. Synthesizing evidence of diagnostic accuracy. Lippincott Williams Wilkins Medical Research, Philadelphia, Pa.

Kadavath Meethal, N.S., Mazumdar, D., Asokan, R., Panday, M., van der Steen, J., Vermeer, K.A., Lemij, H.G., George, R.J., Pel, J.J.M., 2018. Development of a test grid using Eye Movement Perimetry for screening glaucomatous visual field defects. Graefe's Archive for Clinical and Experimental Ophthalmology 256, 371–379.

Kita, Y., Holló, G., Saito, T., Murai, A., Kita, R., Hirakata, A., 2019. Circumpapillary microperimetry to detect glaucoma: a pilot study for sector-based comparison to circumpapillary retinal nerve fiber layer measurement. International Ophthalmology 39, 127–136.

Klein, B.E., Klein, R., Sponsel, W.E., Franke, T., Cantor, L.B., Martone, J., Menage, M.J., 1992. Prevalence of glaucoma. The Beaver Dam Eye Study. Ophthalmology 99, 1499–1504.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lawrenson, J.G., Edgar, D.F., Spry, P.G.D., 2014. Glaucoma – early detection and disease monitoring: an optometric challenge. Optometry in Practice 5, 161–168

Leeflang, M.M., Deeks, J.J., Takwoingi, Y., Macaskill, P., 2013. Cochrane diagnostic test accuracy reviews. Systematic Reviews 2.

Meethal, N.S.K., Pel, J.J.M., Mazumdar, D., Asokan, R., Panday, M., van der Steen, J., George, R., 2019. Eye Movement Perimetry and Frequency Doubling Perimetry: clinical performance and patient preference during glaucoma screening. Graefe's Archive for Clinical and Experimental Ophthalmology 257, 1277–1287.

Michelessi, M., Lucenteforte, E., Miele, A., Oddone, F., Crescioli, G., Fameli, V., Korevaar, D.A., Virgili, G., 2017. Diagnostic accuracy research in glaucoma is still

incompletely reported: An application of Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015. PLOS ONE 12, e0189716.

Mitchell, P., Smith, W., Attebo, K., Healey, P.R., 1996. Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study. Ophthalmology 103, 1661–1669.

Nema, H.V., Nema, N. (Eds.), 2014. Diagnostic procedures in ophthalmology, Third edition. ed. Jaypee Brothers Medical Publishers (P) LTD, New Delhi; Philadelphia.

Olsen, A.S., Cour, M., Damato, B., Kolko, M., 2019. Detection of visual field defects by opticians – with Damato Multifixation Campimetry Online. Acta Ophthalmologica 97, 577–582.

Paranjothy, B., Shunmugam, M., Azuara-Blanco, A., 2007. The Quality of Reporting of Diagnostic Accuracy Studies in Glaucoma Using Scanning Laser Polarimetry: Journal of Glaucoma 16, 670–675.

Quigley, H.A., Vitale, S., 1997. Models of open-angle glaucoma prevalence and incidence in the United States. Investigative Ophthalmology and Visual Science 38, 83–91.

Ramakrishnan, R., Nirmalan, P.K., Krishnadas, R., Thulasiraj, R.D., Tielsch, J.M., Katz, J., Friedman, D.S., Robin, A.L., 2003. Glaucoma in a rural population of southern India: the Aravind comprehensive eye survey. Ophthalmology 110, 1484–1490.

Roever, L., 2016. Types of Bias in Studies of Diagnostic Test Accuracy. Evidence based Medicine and Practice 2.

Shunmugam, M., Azuara-Blanco, A., 2006. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. Investigative Ophthalmology and Visual Science 47, 2317–2323.

Siddiqui, M. a. R., Azuara-Blanco, A., Burr, J., 2005. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. British Journal of Ophthalmology 89, 261–265.

Šimundić, A.-M., 2009. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC 19, 203–211.

Thapa, S.S., Paudyal, I., Khanal, S., Twyana, S.N., Paudyal, G., Gurung, R., Ruit, S., van Rens, G.H.M.B., 2012. A population-based survey of the prevalence and types of glaucoma in Nepal: the Bhaktapur Glaucoma Study. Ophthalmology 119, 759–764.

The EQUATOR Network. 2018, http://www.equator-network.org/, accessed 08/01/20.

Tielsch, J.M., Katz, J., Singh, K., Quigley, H.A., Gottsch, J.D., Javitt, J., Sommer, A., 1991. A population-based evaluation of glaucoma screening: the Baltimore Eye Survey. American Journal of Epidemiology 134, 1102–1110.

Vijaya, L., George, R., Baskaran, M., Arvind, H., Raju, P., Ramesh, S.V., Kumaramanickavel, G., McCarty, C., 2008. Prevalence of primary open-angle glaucoma in an urban south Indian population and comparison with a rural population. The Chennai Glaucoma Study. Ophthalmology 115, 648-654.e1.

Weih, L.M., Nanjan, M., McCarty, C.A., Taylor, H.R., 2001. Prevalence and predictors of open-angle glaucoma: results from the visual impairment project. Ophthalmology 108, 1966–1972.

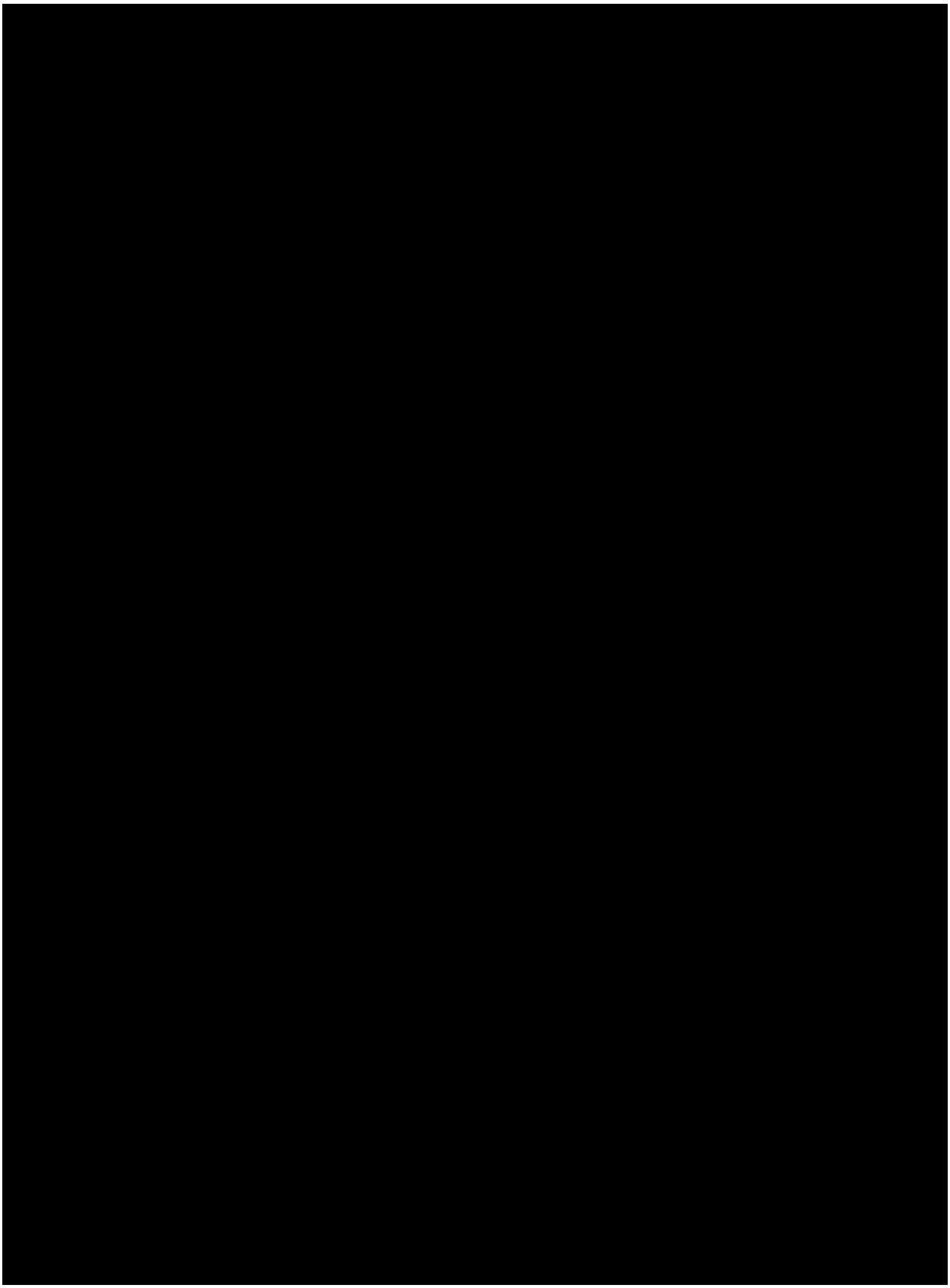
Whiting, P., Rutjes, A.W., Reitsma, J.B., Bossuyt, P.M., Kleijnen, J., 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Medical Research Methodology 3.

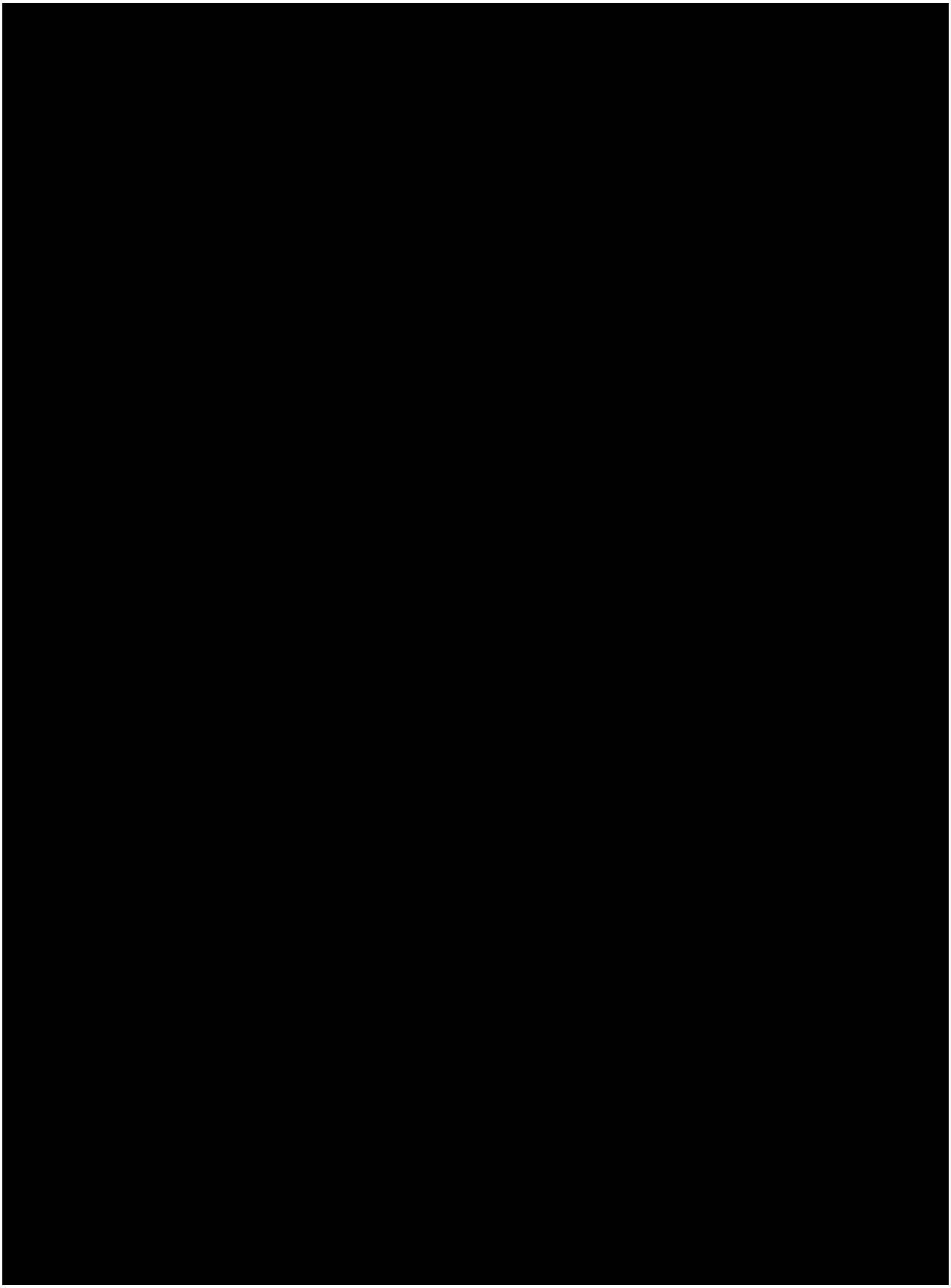
Whiting, P.F., 2011. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Annals of Internal Medicine 155, 529.

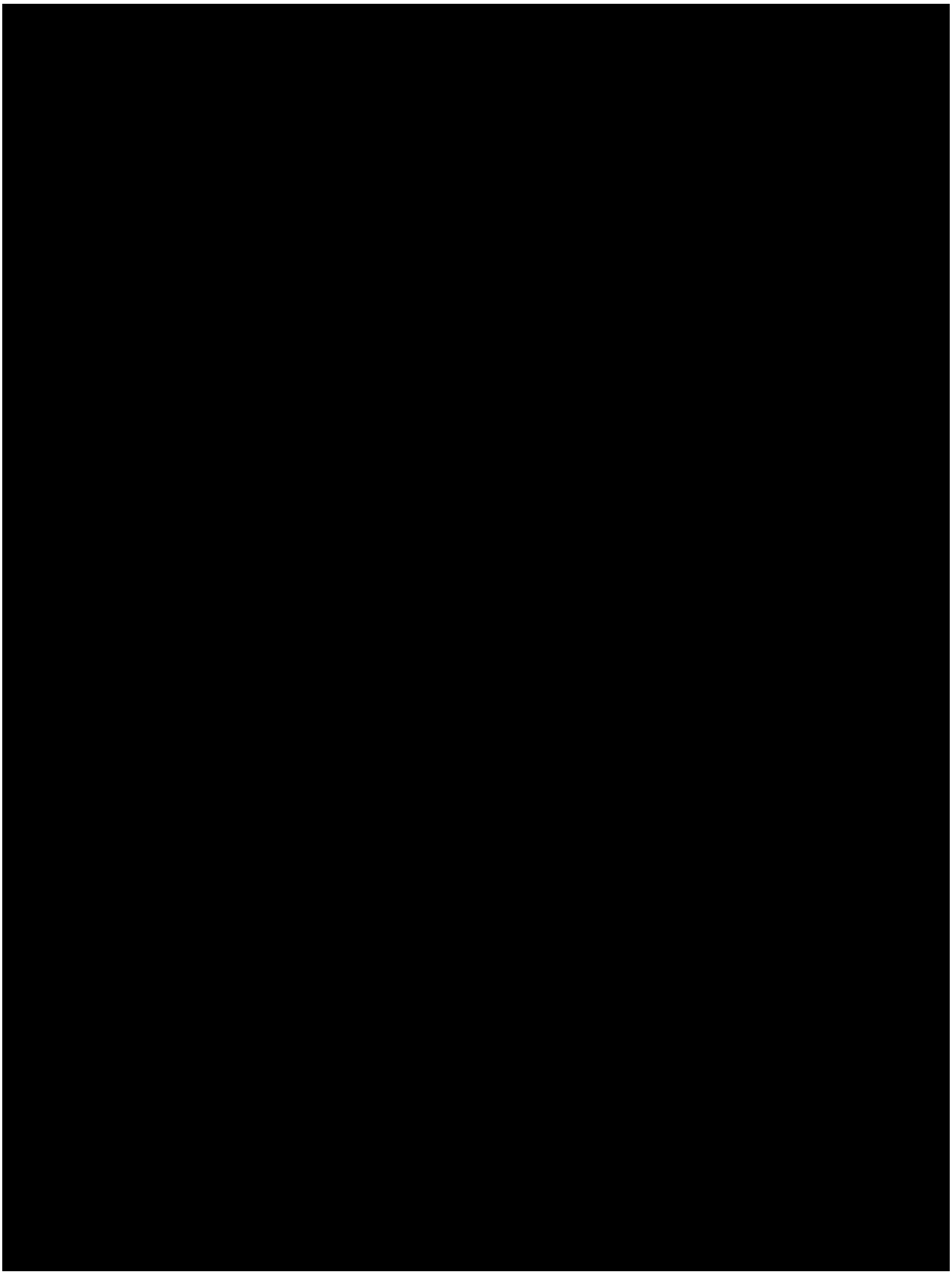
Wilczynski, N.L., 2008. Quality of Reporting of Diagnostic Accuracy Studies: No Change Since STARD Statement Publication—Before-and-after Study. Radiology 248, 817–823.

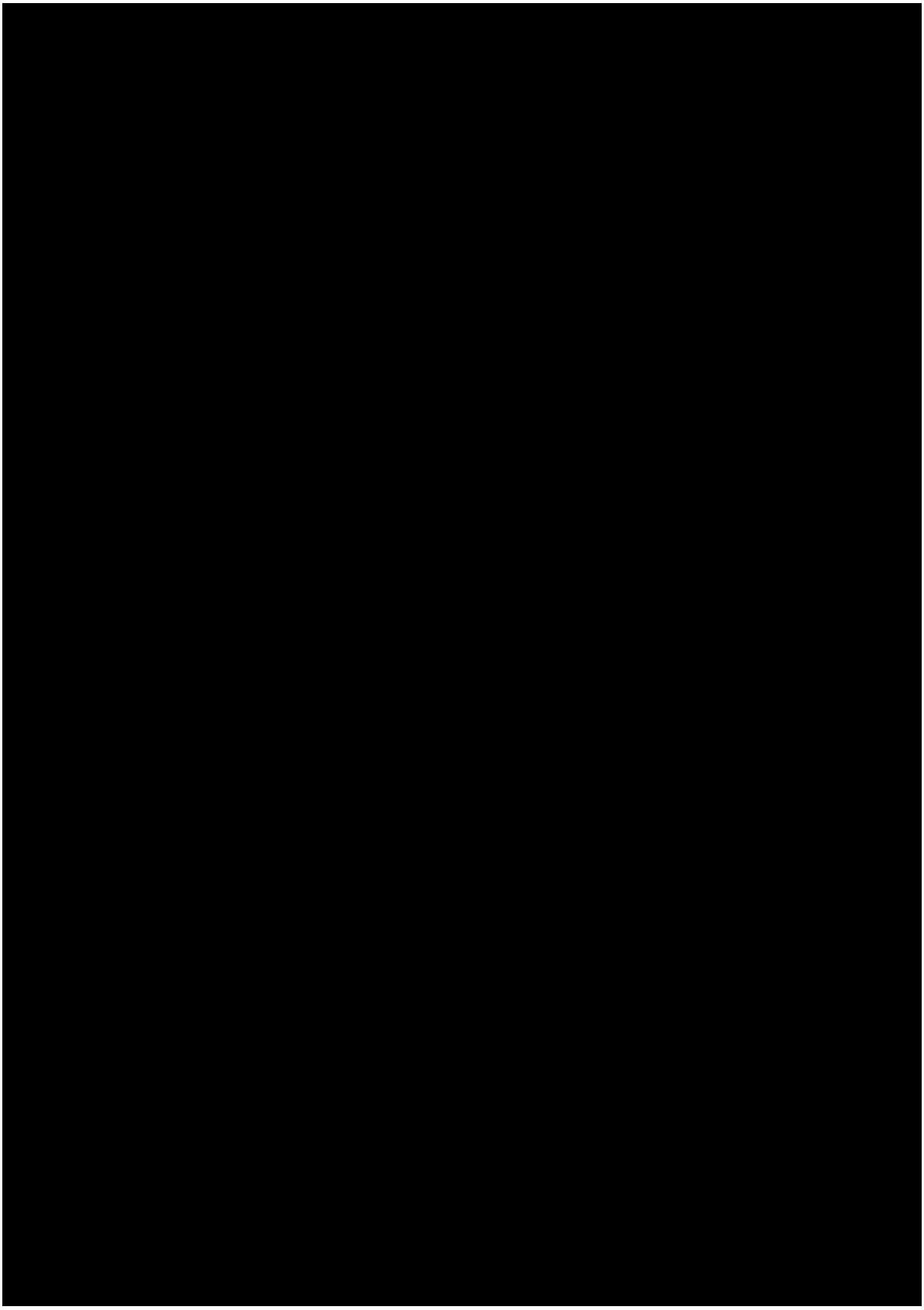
Zafar, A., Khan, G.I., Siddiqui, M.A.R., 2008. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: a systematic review. Clinical and Experimental Ophthalmology 36, 537–542.

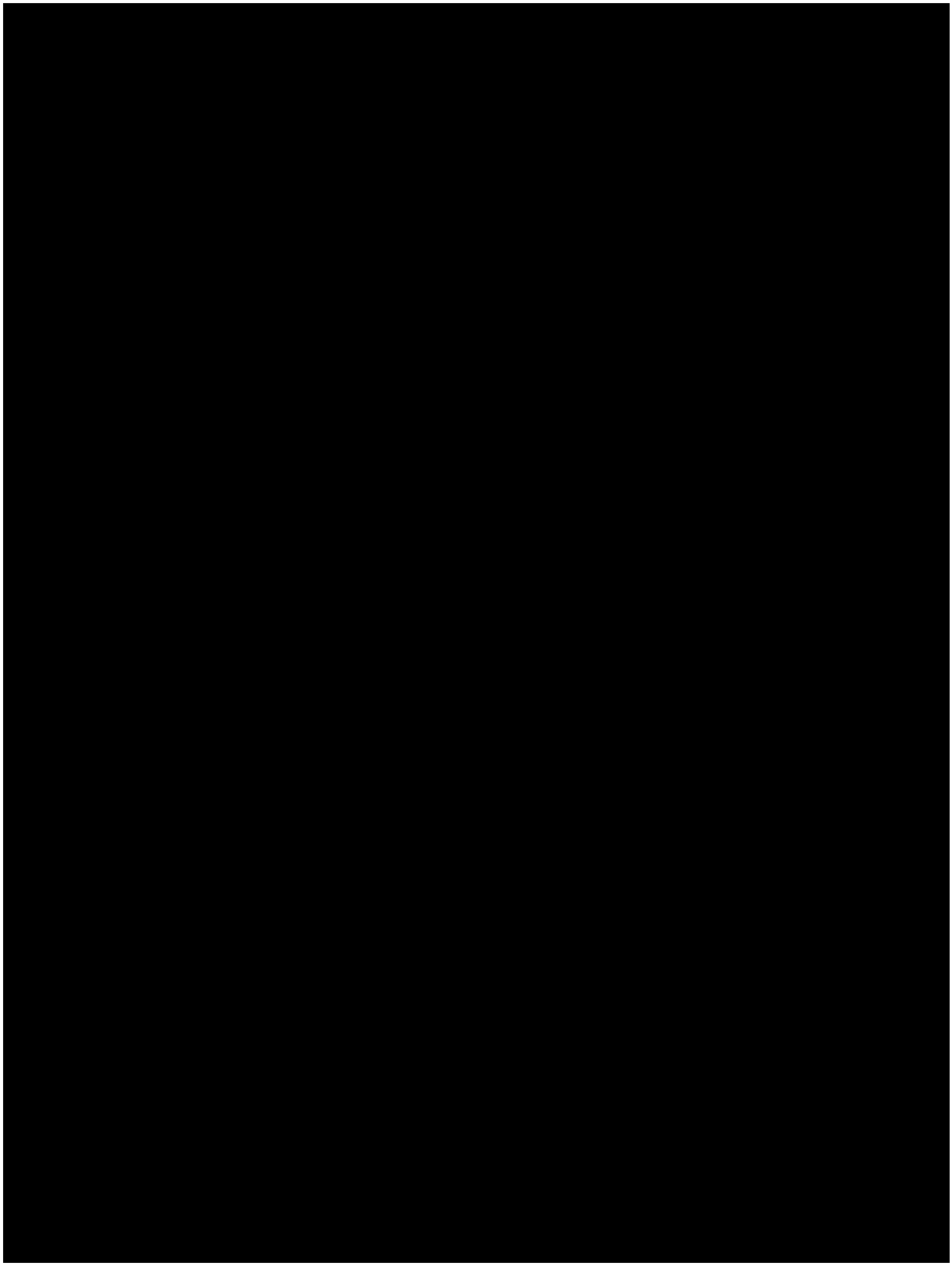
2.7. Published Article

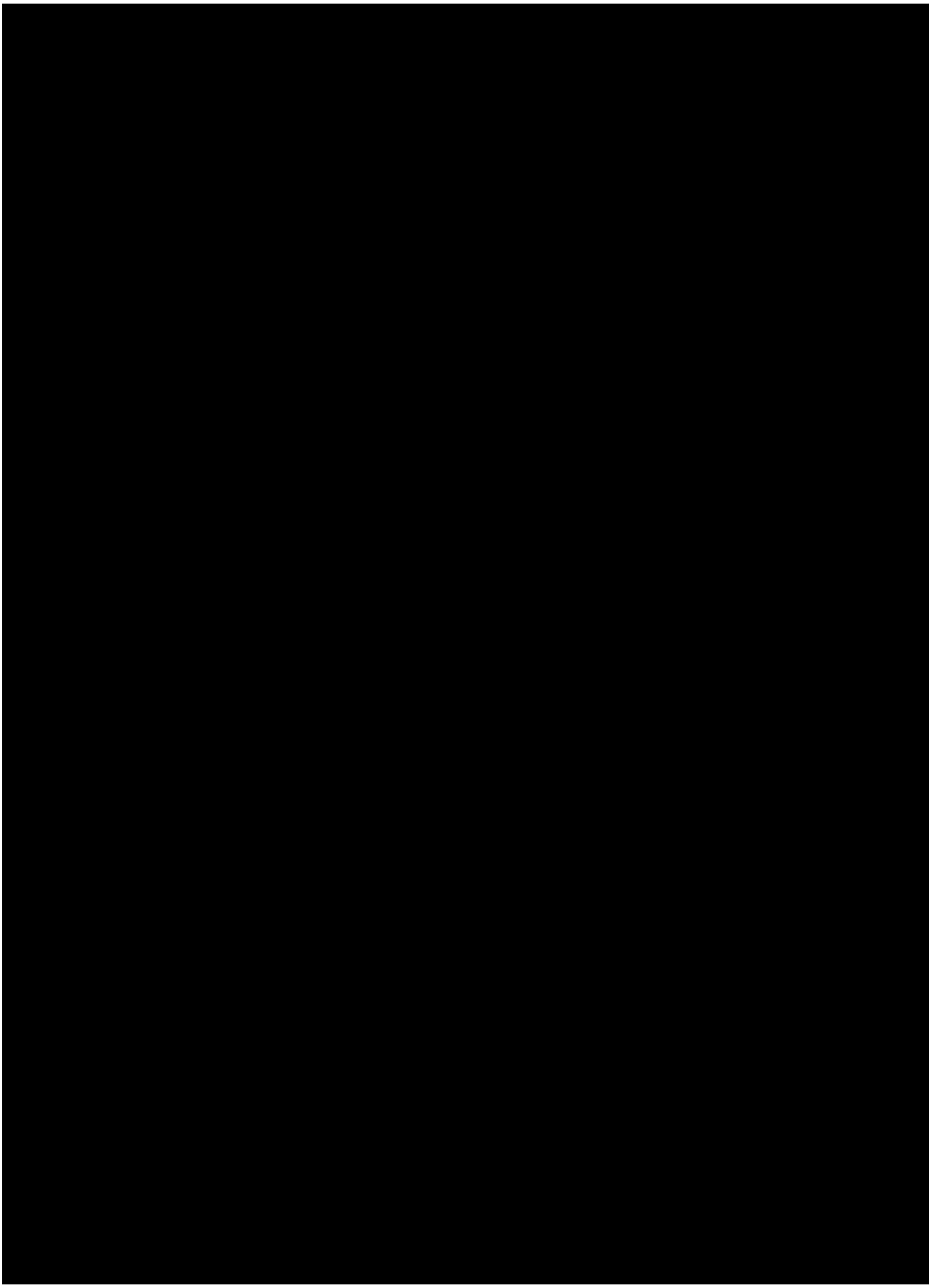


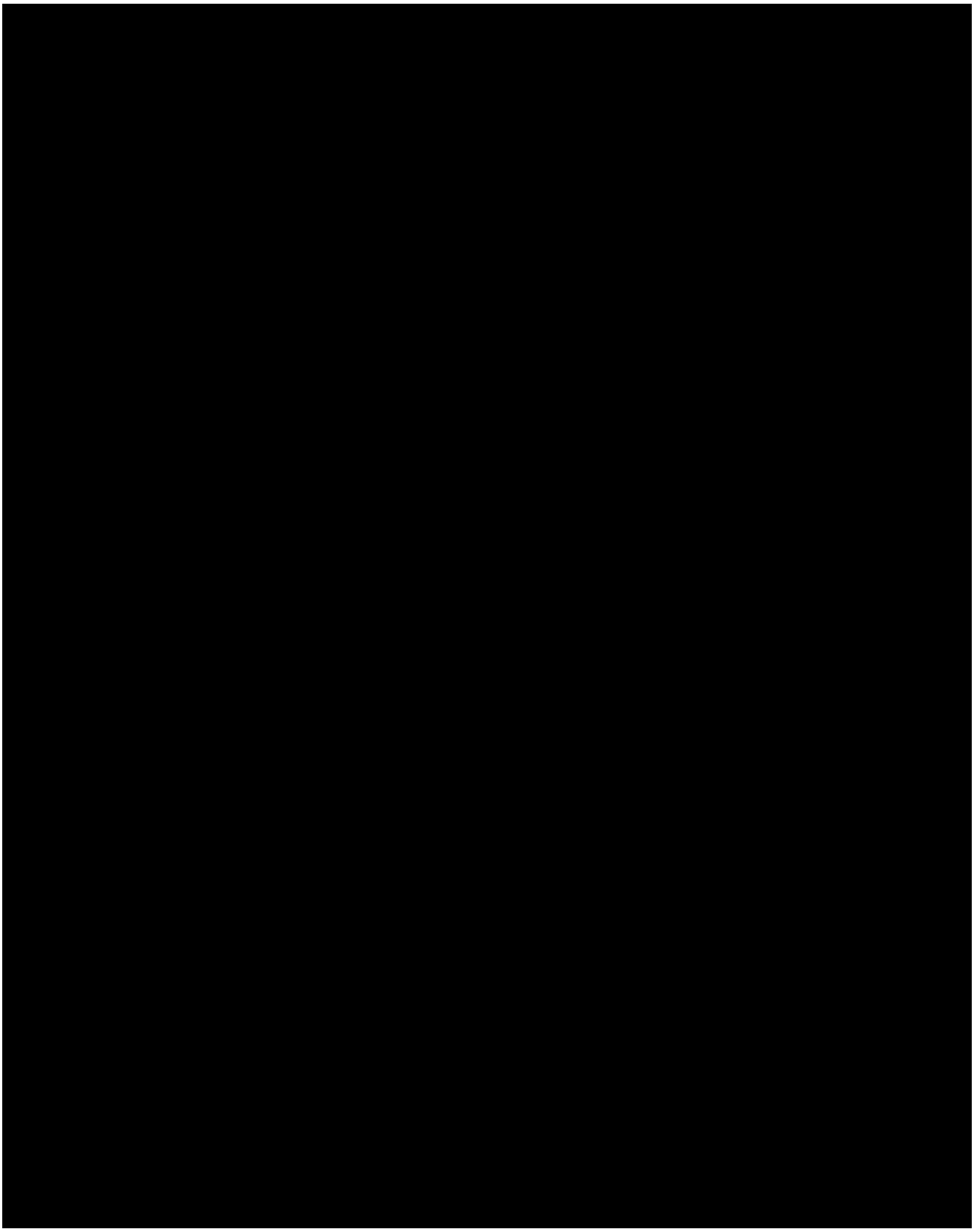


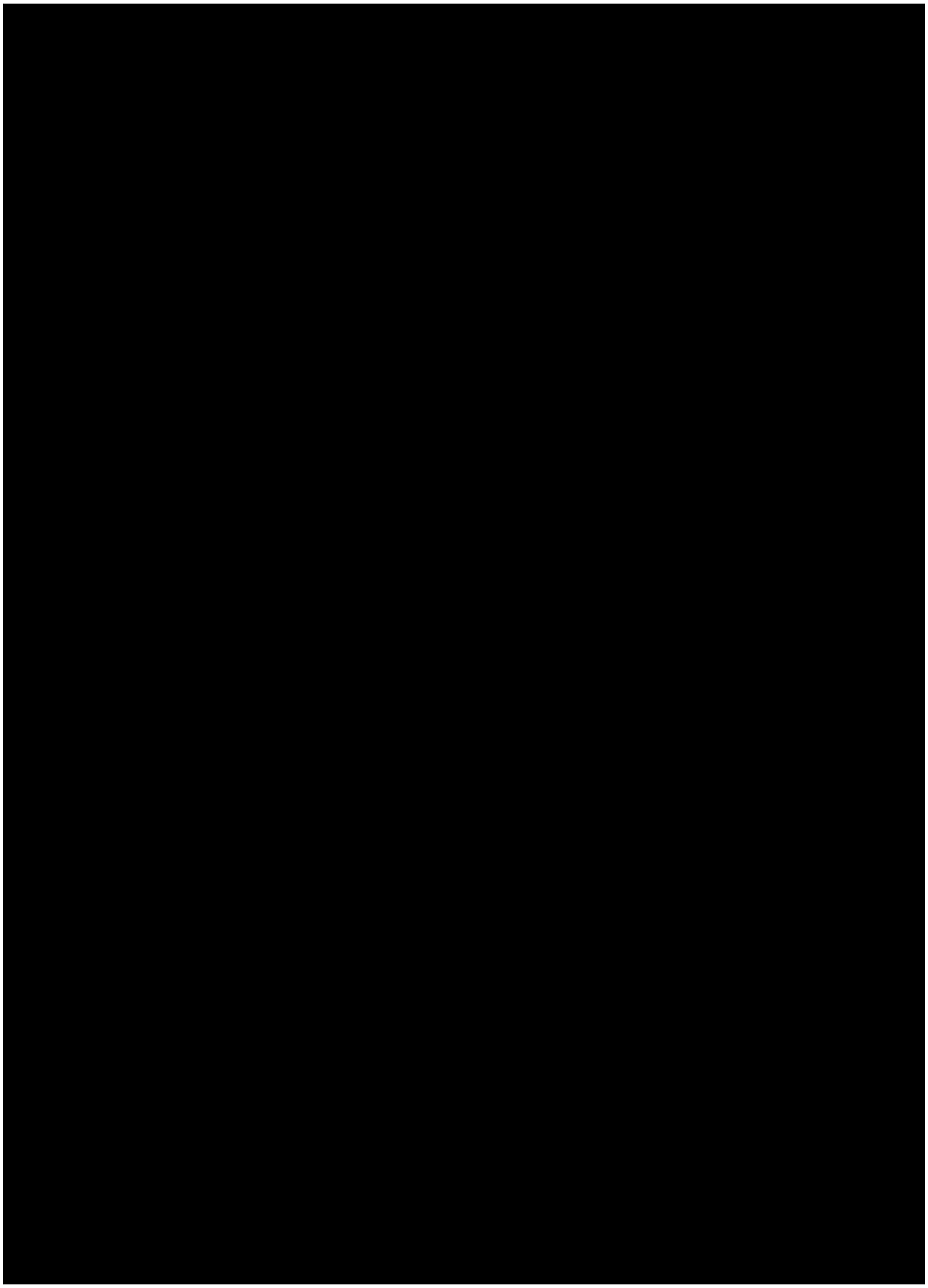


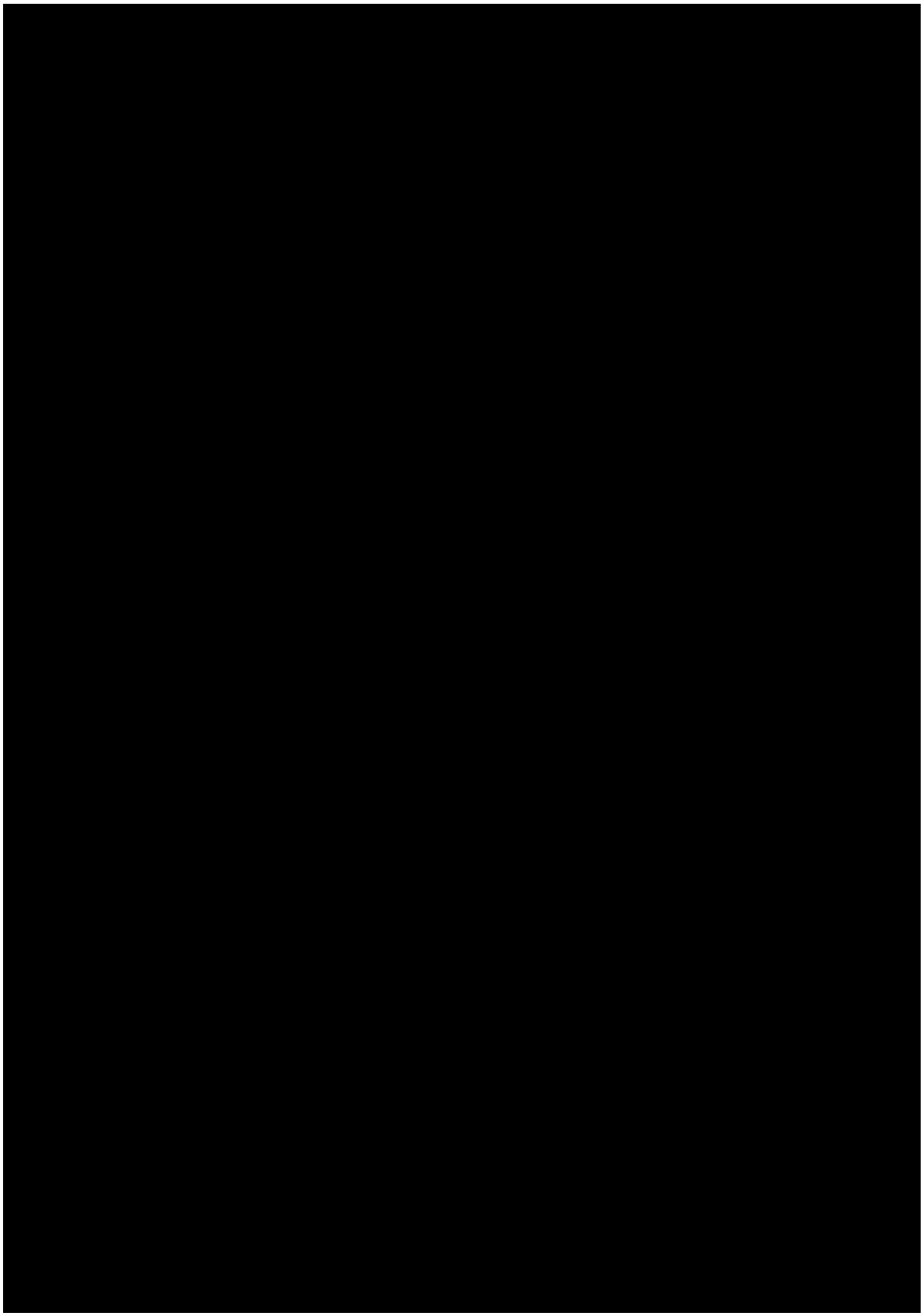












Chapter 3.

Development of a new glaucoma screening test using temporally modulated flicker

3.1 Introduction

Glaucoma is a leading cause of visual morbidity, accounting for 6.6% of blindness globally (Bourne et al 2013). As the disease is typically asymptomatic in its early stages, many patients already have significant functional visual loss at the time of diagnosis. Although the disease fulfils many of the Wilson-Jungner criteria (Wilson et al 1968) to justify the development of a screening programme, screening of the general population for glaucoma has not been found to be cost-effective in any country (Hernandez et al 2008), although there is the potential for targeted screening of high-risk groups (e.g. those of African ancestry or with a family history of glaucoma) (Burr et al 2007). As an alternative to the current practice of opportunistic screening (case-finding) for chronic open angle glaucoma (COAG) (Figure 3.1A), a pathway has been proposed that could potentially improve cost-effectiveness by performing an initial technology-based assessment to determine the probability of disease and then referring those who require a more detailed ophthalmic assessment (Figure 3.1B) (Burr et al 2007).

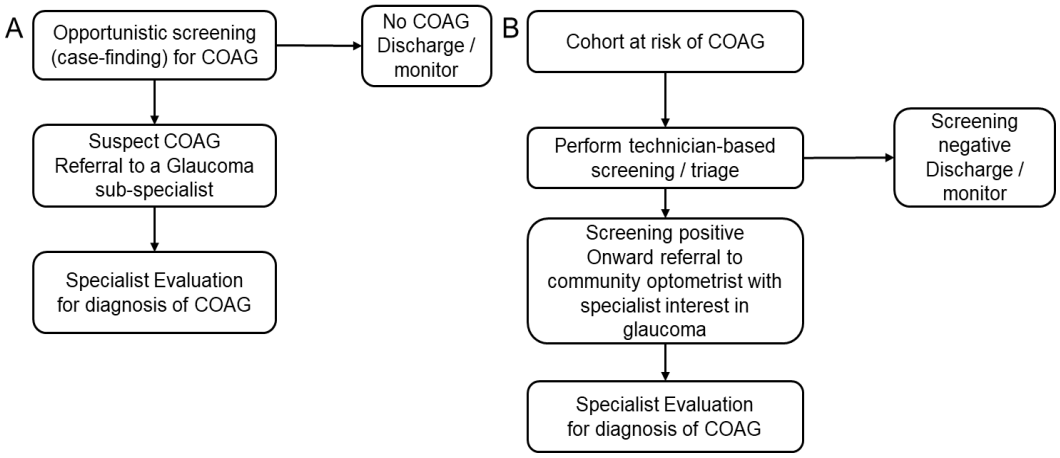


Figure 3.1. – A). Current opportunistic screening pathway for COAG. B). Alternative pathway incorporating a technology-based triage assessment. To be consistent with NICE guidance, specialist evaluation for the definitive diagnosis of COAG is required by a consultant ophthalmologist based on gonioscopy, threshold central automated perimetry, IOP using Goldmann applanation tonometry and optic nerve assessment.

A systematic review evaluating screening tests for detecting COAG concluded that no test or group of tests is clearly superior for glaucoma screening (Mowatt et al 2008). As well as being able to accurately distinguish normal individuals from those who have the disease, an effective screening test for glaucoma should ideally be easy to administer and interpret, portable, quick, and be acceptable to the people being tested. Historically, the reference standard for assessing visual function in glaucoma has been standard automated perimetry (SAP). However, SAP has a number of shortcomings: the test is time consuming, it lacks portability, and the technique is associated with high intra-participant variability (Spry et al 2001). Given the limitations of SAP, several studies have investigated the value of alternative psychophysical tests (McKendrick et al 2005).

Reduced sensitivity to temporally modulated sinusoidal flicker has been shown to provide an indicator of compromised retinal ganglion cell function, suggesting that assessment of flicker sensitivity could be a potentially useful method for detecting glaucomatous damage (Tyler et al 1981; Lachenmayr et al 1992; Horn et al 1997). Contrast modulation flicker uses a stimulus that is matched in luminance to the background. The contrast of the stimulus is then modulated temporally according to a fixed frequency, and the amplitude of flicker modulation needed for detection of the stimulus is determined (Tyler et al 1991). When presenting this stimulus, it is important to avoid sudden stimulus onsets and offsets that can disrupt the ability to detect the flicker at a particular frequency. To prevent this, flicker modulation stimuli are usually presented within a temporal cosine envelope or Hanning window (Figure 3.2).

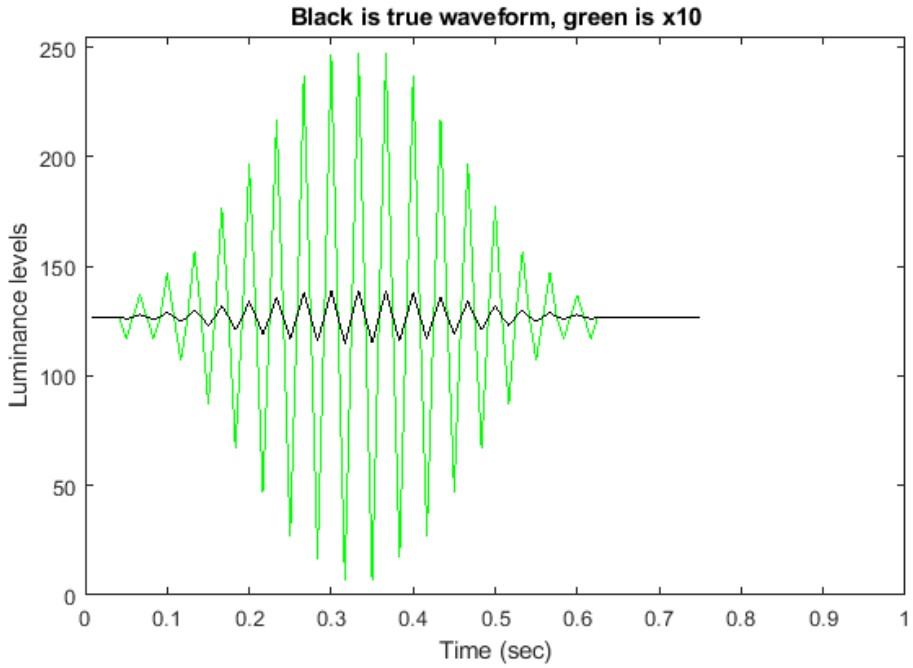


Figure 3.2. – Flicker modulation stimulus within cosine envelope

The aim of the present study was to develop a new psychophysical algorithm to determine flicker sensitivity thresholds in susceptible areas of the visual field that could be used as a rapid screening test for COAG. Given the low prevalence of the disease in the general population, a large proportion of those screened would be expected to be normal or close to normal, therefore it is important to initially assess the performance of the algorithm in a disease-free population, to determine how efficiently they could be screened and gain an initial estimate of the stability, test-retest reliability, speed and physiological variation in flicker sensitivity in the normal population that would define the specificity of the test.

3.2. Materials and methods

This study was approved by the School of Health Sciences Research Ethics Committee, City, University of London and conducted according to the tenets of Declaration of Helsinki. Written informed consent was obtained from all participants.

Development of the Accelerator 4 alternative forced choice flicker test prototype

Stimulus Design

Previous research has suggested that with comparatively few perimetric stimuli, it is possible to achieve high sensitivity for glaucoma detection (Henson et al 1988; Krakau et al 1989; De la Rosa et al 1990; Sugimoto et al 1998), which reflects the strong inter-point correlations within regions of the visual field that are adversely affected in glaucoma. The Accelerator 4 alternative forced choice flicker test prototype (A4FTp) was designed to combine this approach with the efficiency of a four-alternative forced-choice strategy for psychophysical testing. The test involves the participant undertaking a series of trials to detect a target stimulus from four possible stimulus locations (Figure 3.3). The selection of stimulus locations for this first iteration of the A4FTp was based on the research by Wang and Henson (Wang et al 2013), who used optimized sub-sets of the conventional 24-2 test pattern based on the positive predictive value (PPV) of each test location to identify glaucomatous visual field loss (Nicholas et al 1980; Keltner et al 2003). The A4FTp uses two 11° diameter circular stimuli located in the temporal superior and inferior arcuate regions of the visual field 9–21° from fixation (Figure 3.3), with the other two equivalently sized stimuli spanning the horizontal meridian (14–26° from fixation), corresponding to the location of the 'nasal step' that is often seen in COAG. The configuration was mirror reversed for testing the left eye.

The target stimulus was a 0.75 s period of 30 Hz sinusoidal flicker ramped on and off according to a raised cosine envelope in order to avoid onset and offset transients, presented at a viewing distance of 33 cm on a high refresh rate screen (120 Hz) in a uniform red field (610 nm) with a mean background luminance of 19 cd m^{-2}. The presentation with long-wavelength light was designed to minimise transmission losses in the optic media.

To ensure that the luminance of the stimulus is matched to the luminance of the background, the display gamma, a function that controls overall brightness of an image, was linearized to within 0.5% of the maximum intensity to avoid any change in luminance as the flicker amplitude is modulated.

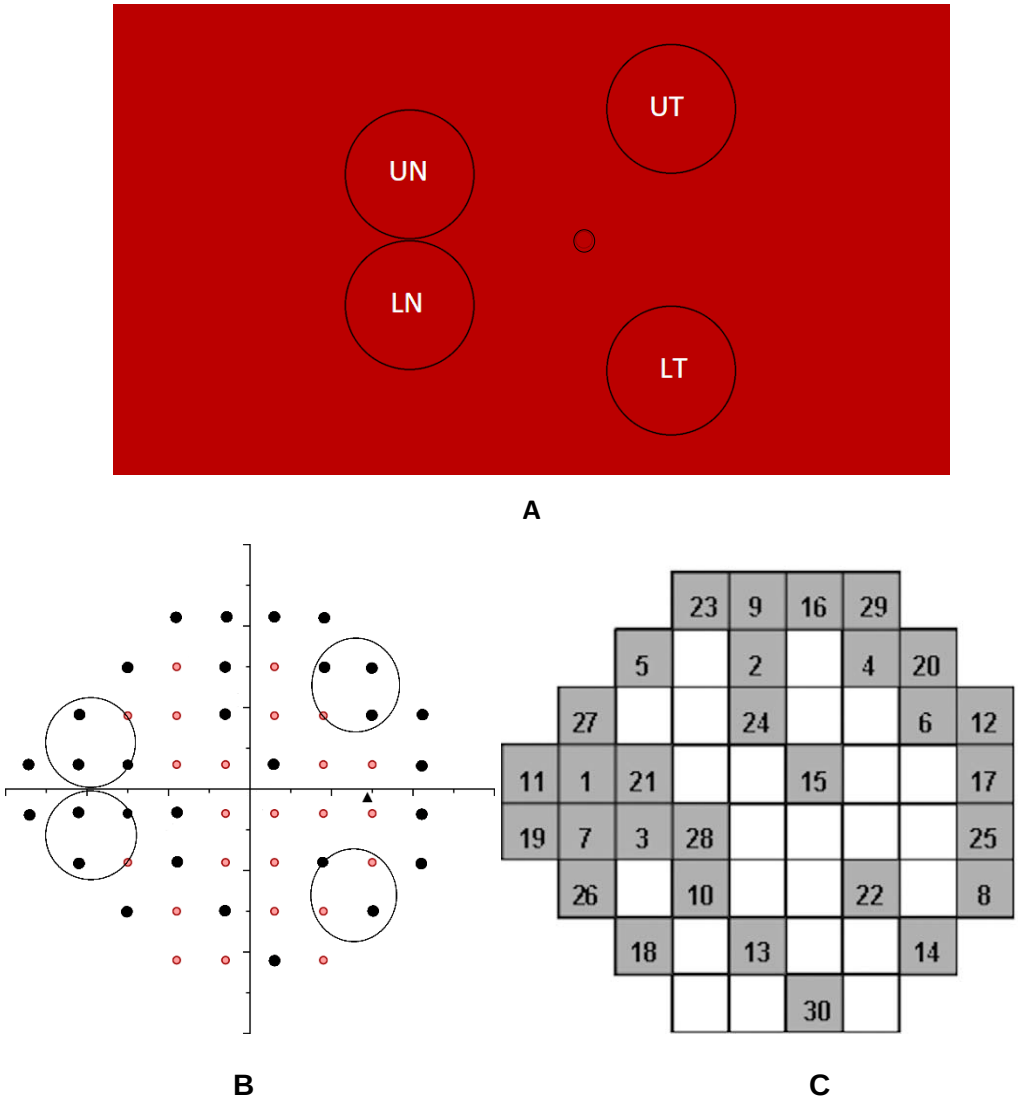


Figure 3.3. A. Right-eye spatial arrangement for the A4FTp test stimuli presented in a red field on a high refresh rate screen. The mirror image configuration was used for left-eye viewing. **B.** Spatial location of the A4FTP flicker test locations for the right eye compared to the HFA 24-2 pattern. **C.** Field map by Wang and Henson (Wang et al, 2013). The configuration was mirror reversed for testing the left eye. (UN – Upper Nasal, LN – Lower Nasal, UT- Upper Temporal, LT – Lower Temporal)

Algorithm development

The glaucoma screening algorithm uses a rapid four-alternative forced-choice staircase paradigm that we term the 'Accelerator method' to measure threshold sensitivity for 30 Hz sinusoidal flicker simultaneously in four peripheral field regions by means of a novel staircase termination criterion, such that the one-up/two-down staircase in log modulation steps of 1 dL terminates when the standard deviation of the last n trials becomes less than 1 step. Unlike conventional staircases, this straightforward algorithm ensures that the staircase reaches a stable asymptotic performance level of low variability before terminating.

The performance of the staircase depends on the value chosen for n, and the steps should be defined in a domain where the variance is uniform. Since most psychophysical performance operates in domains where the variance is proportional to signal strength (Weber's Law), it is generally best to run the staircase in logarithmic steps. Since, it is helpful to have a definitive change in the stimulus on each trial, steps of 1 dL (0.1 $\log_{10}$ units) are an effective choice for the step size.

The algorithm was implemented using a Dell computer (Inter core 2 Duo CPU E7500 @2.93 GHz and 4 GB RAM), displayed on a 144 Hz Asus VG248QE 24-inch 3D LED monitor running at 120 Hz, and used an Accmat™ USB wired 19 key number numeric keypad as the test input device. The computer was running on a 64 - Bit Windows 7 Enterprise, Service pack 1, MATLAB R2014a and Psychtoolbox version 3.0.11.

The code for the A4FTp flicker algorithm is provided in Appendix 2.

Algorithm performance simulation

To illustrate the performance of the A4FTp staircase method, simulations of its performance were run using Gaussian noise with a standard deviation (sigma) of 0.5, 1 or 2 steps of the staircase, with 100 runs under each condition (Figure 3.4).

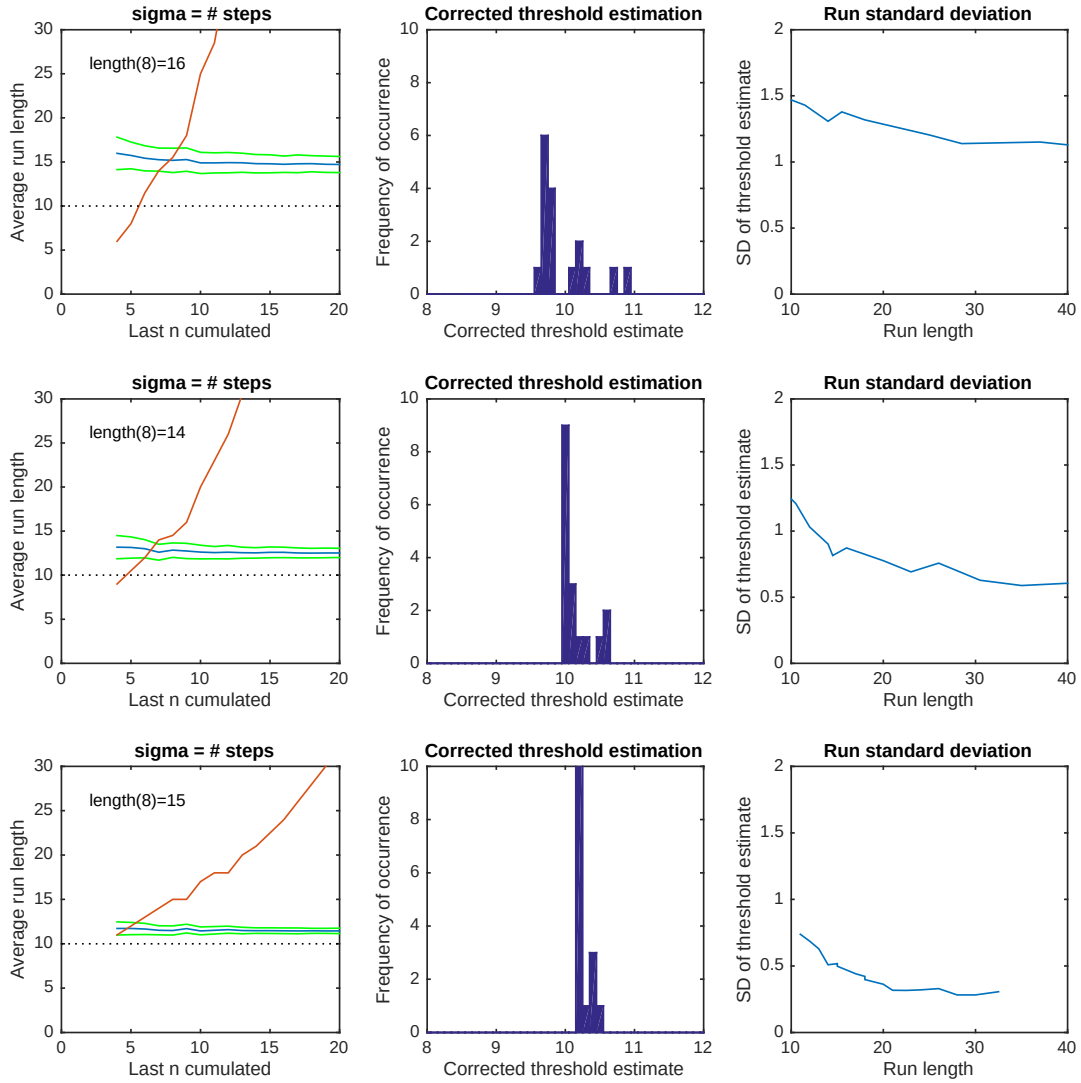


Figure 3.4. Simulation of the performance of the Accelerator staircase method at three levels of step size relative to the noise sigma controlling threshold level (0.5, 1 and 2 decilogs top to bottom row respectively). Column 1: Overlays of average run length (red curves) and threshold estimate (blue curves) ± 1 standard error of the mean (SEM) (green curves) as a function of the criterion cumulation length for the last n values for the sigma estimate. Dashed line is the simulated threshold level. Inset gives the average run length for a criterion cumulation length of 8 steps. Column 2: Histograms of the corrected threshold estimates on each trial simulation. Column 3: Standard deviation (SD) of the threshold estimates as a function of run length. Simulation provided by Christopher Tyler.

The left column shows the average threshold estimates (blue curves) ± 1 standard error of the mean (SEM, green curves) as a function of number of trials cumulated (n). The average is almost independent of the n. It also plots the average run length as a function of the criterion run length (red curves). As expected, the average run length increases with n, but it can be seen to do so more steeply when the step size is small relative to the experimental noise sigma than when it is large (from upper left to lower left). For small n, however, the run length remains independent of step size, while standard deviation (SD) decreases with step size, so a larger step size is to be preferred. Standard deviation decreases with run length substantially up to about run lengths of 20 (termination criteria of about 14), to about 0.2 sigma for a step size of 2 and about 0.3 sigma for a step size of 1, again favouring the larger step sizes.

The other feature of the average threshold estimates (blue curves in the left column of Figure 3,3) is that they lie consistently above the simulated threshold level by an amount that is almost invariant with n but is inversely related to step size. The resultant threshold estimates (other columns) were therefore corrected for the overestimation of the raw threshold levels θ_{RAW} seen in the left column according to the empirical formula:

$$\theta_{CORRECTED} = -a \,.\, \theta_{RAW} \,/\, k, \qquad \text{where } a = 1 \text{ dL.}$$

The middle column shows histograms of the corrected threshold estimates resulting from 100 simulations for each of the three step sizes, k. Note that the histograms have a narrow spread, decreasing with step size, and that the peak of the corrected estimates is close to the defined threshold level of 10.

The right column shows the average simulated SD as a function of run length for the three step sizes, asymptoting to much less than 1 dL in all cases. Note that there is little advantage for run lengths longer than 14.

The data obtained with the simulation allowed us to select the 8 step (T8) and 12 step (T12) termination criterion used for the subsequent experimental studies.

Testing protocol

A convenience sample of normal participants was recruited from City, University of London staff and students. To be eligible for the study, participants were required to be 18 years or older with a normal appearance of the optic disc and fundus and no known family history of glaucoma in first-degree relatives. All volunteers underwent a detailed ophthalmic examination, including SAP (HFA 24-2 SITA Standard), to exclude glaucoma or any other ocular diseases that would affect flicker sensitivity. Participants were excluded if they had an intraocular pressure (IOP) greater than 21 mm Hg in either eye, suspicious appearance of the optic disc (rim loss, optic disc haemorrhage, etc.), history of ocular disease, surgery or trauma, history of a cerebrovascular event, or diabetes mellitus. None of the participants had previous experience of flicker perimetry.

Selection of suitable threshold criteria

An initial study was carried out to determine the optimal threshold termination criteria (by comparing the performance of the T8 and T12 run lengths) and to assess any learning effects that could potentially influence the percentage modulation levels. For 20 healthy adults, one eye selected at random underwent testing with the flicker test four times on separate occasions during a period of two weeks for each criterion run length.

Participants were given an identical set of instructions on how to perform the test. These consisted of:

- Explaining the general layout of the screen and the number, position and timings of the stimuli
- Familiarising the participant with the numeric keypad and the relationship between the keys and stimulus location on the screen
- Explaining the need to fixate centrally and to select the key on the keypad corresponding to the location of each presented stimulus immediately after the auditory signal
- To press any key if unsure of the stimulus location or if no stimulus was seen

Participants wore near reading prescription if needed, and the eye not being tested was occluded with an eye pad. A practice run was provided prior to performing the actual test

The starting level was set at 15 dL, which for healthy individuals, is halfway between the threshold of about 10 dL and the maximum modulation level of 20 dL. For each session, the overall mean modulation level was determined for the four test locations and the time taken to conduct the test was recorded. Inter-session variations were assessed using the Friedman Test and test-retest coefficients of repeatability for the flicker modulation levels were calculated.

In order to evaluate the inherent variability of the test for each participant, the inter-participant variability was isolated from the intra-participant variability by calculating the mean modulation level across participants for each test location and each threshold criterion and then normalizing the individual values to this group mean to determine the within-participant variability.

Assessment of individual reliability and learning effects

To study individual reliability, we evaluated the test-retest repeatability of the flicker algorithm. First, we selected 4 volunteers to assess the individual reliability and any tendency to improve by learning by taking the test 10 times over a period of 3 months. Learning effects were assessed by comparing the results of test duration and modulation levels of the first session with those of the other four sessions using either Friedman's test or Wilcoxon's test as appropriate. The intra-participant test–retest coefficients of repeatability (CoR) were also calculated for the flicker modulation levels. All data and statistical analyses were performed using Excel 2007 and SPSS version 20.

3.3. Results

Determination of optimal threshold criteria

Twenty normal participants meeting our inclusion criteria were included in the study. Seventy per cent were male, mean age 33.8 years (SD 8.5), mean spherical refractive error -0.50 ±2.41D. The analysis was based on 10 right eyes and 10 left eyes.

Table 3.1 shows the average inter-session results for each threshold criterion (T8 vs T12). A non-parametric Friedman test for the modulation levels rendered a Chi-square of 6.00, $p = 0.111$ for the shorter termination criterion (T8) and a Chi-square of 5.99, $p = 0.112$ for the longer termination criterion (T12). The Friedman test for run duration rendered a Chi-square of 3.214, $p = 0.36$ for the shorter T8 termination criterion and 17.82, $p < 0.001$ for the longer T12 criterion. Moreover, for this larger sample the mean durations were nearly twice as long for the T12 than the T8 criterion.

Table 3.1. Inter-session modulation levels and test durations (per eye)

	Session 1	Session 2	Session 3	Session 4	p-value
% Modulation (T8)	9.22 ± 0.17	9.57 ± 0.27	9.17 ± 0.15	9.07 ± 0.18	0.111
% Modulation (T12)	8.61 ± 0.43	8.69 ± 0.41	8.76 ± 0.40	8.90 ± 0.37	0.112
Duration (T8) (s)	81 ± 22	71 ± 9	76 ± 15	73 ±13	0.36
Duration (T12) (s)	166 ± 49	148 ± 40	148 ± 40	126 ± 31	< 0.001

Figure 3.5 shows the modulation sensitivity levels and 95% confidence intervals in all 4 sessions for both threshold criteria (T8 and T12) respectively.

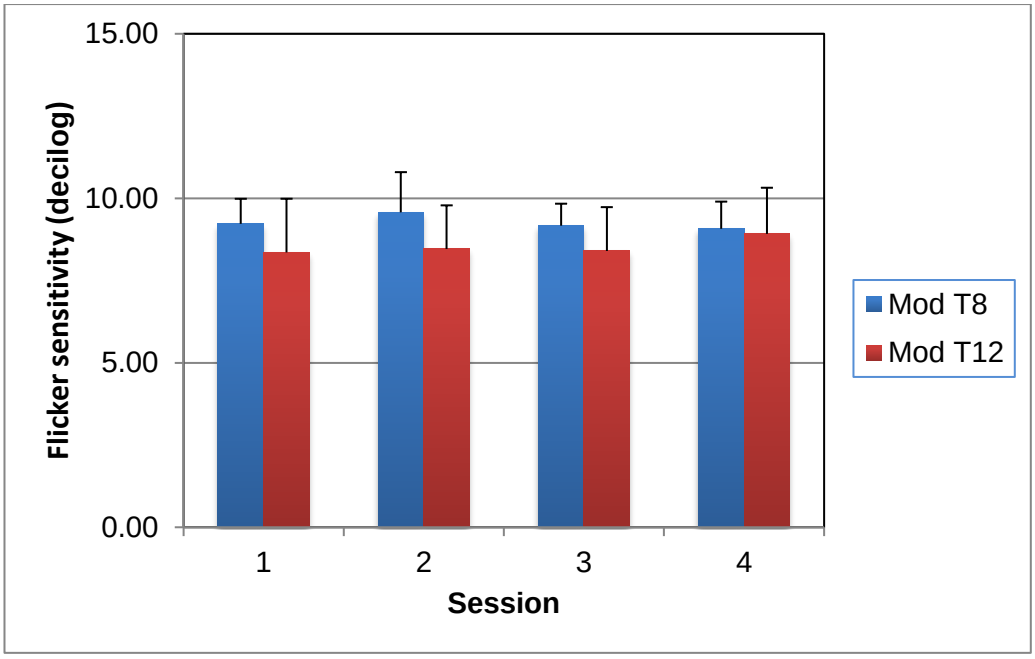
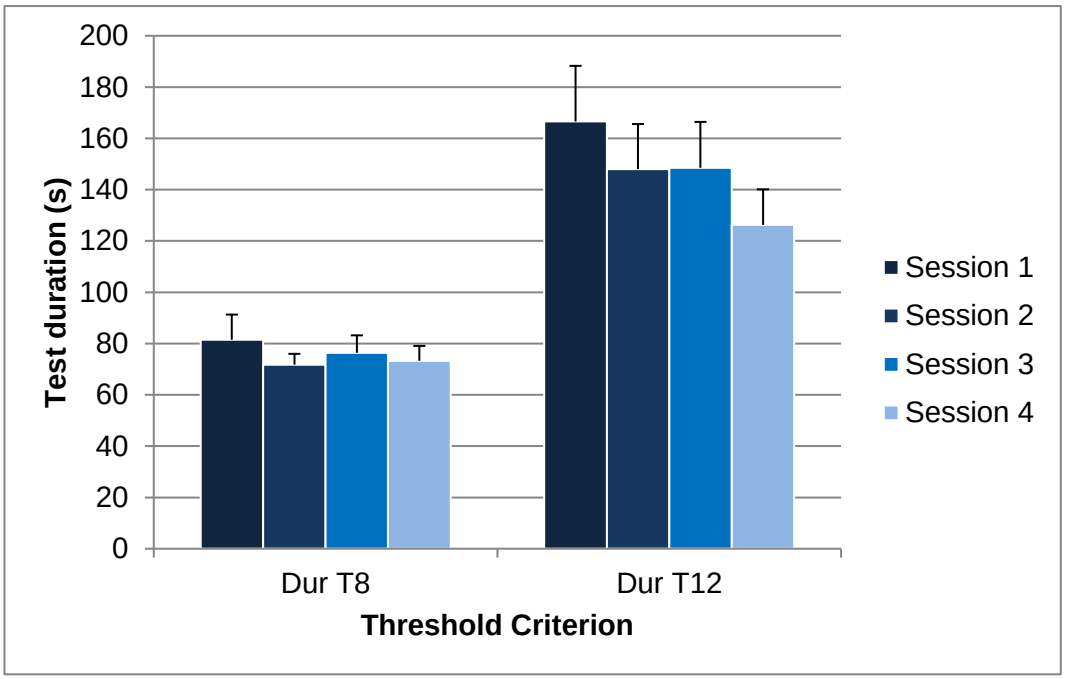


Figure 3.5. Decilog modulation (Mod) levels ±1SD across the 4 sessions for each threshold criterion.



x

Figure 3.6. Test duration (Dur) per eye in seconds ± 1SD across all 4 sessions for each threshold criterion

We can ask whether the threshold values improved over the four test sessions for each participant. The trend is not significant at T8 ($p = 0.25$) but there was a significant decrease in average trial duration for T12 ($p < 0.01$). This result suggests that, in addition to the longer time to reach the criterion at T12, the participants had difficulty managing the stability of their responses for this criterion, while reaching stable performance within the first run for the T12 criterion.

Inherent Variability

To assess the inherent variability of the test for a typical participant, the inter-participant variability needs to be isolated from the intra-participant variability. The global mean sensitivity was thus calculated across participants for each location and each threshold criterion and the individual participant values were then normalised to this group mean for calculation of the within-participant error terms. This procedure removes the across-subject variability without affecting the group mean values.

The results of this analysis are shown in Figure 3.7. There was no significant difference in sensitivity among the four chosen field locations for the group of 20 participants, with either the T8 or T12 criterion. The average within-participant standard deviation across the four locations was 0.52 dL for T8 and 1.32 dL for T12 (thus giving 95% confidence intervals of $\pm$1.01 dL for T8 and $\pm$2.59 dL for T12). The corresponding values for the Bland-Altman Coefficient of Repeatability are 1.44 dL and 3.65 dL). The T8 criterion therefore provides a significant advantage over the T12, even though the T12 was hypothesised to provide a lower variability by enforcing tighter limits.

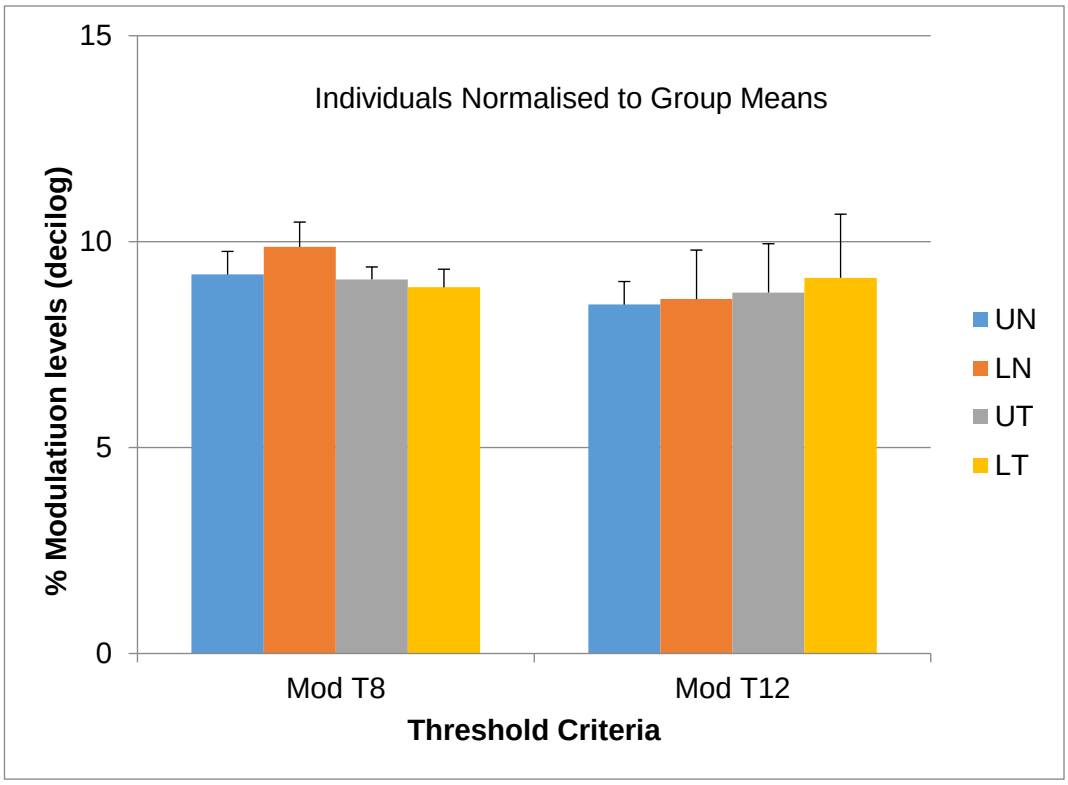


Figure 3.7. Decilog modulation (Mod) levels ±1SD over the four visual field locations, normalised to the group mean, for each threshold criterion. These error bars thus represent the average individual variability at the four visual field locations. (UN – Upper Nasal, LN – Lower Nasal, UT- Upper Temporal and LT – Lower Temporal)

Assessment of individual reliability and learning effects

To study individual reliability, we evaluated the test-retest repeatability of the flicker algorithm. The test was repeated 10 times in four normal observers, shown in Figure 3.8 as the average thresholds across the four locations for each observer. The goal was to test the stability of the algorithm under optimal conditions rather than the performance of typical glaucoma suspects in a clinical setting.

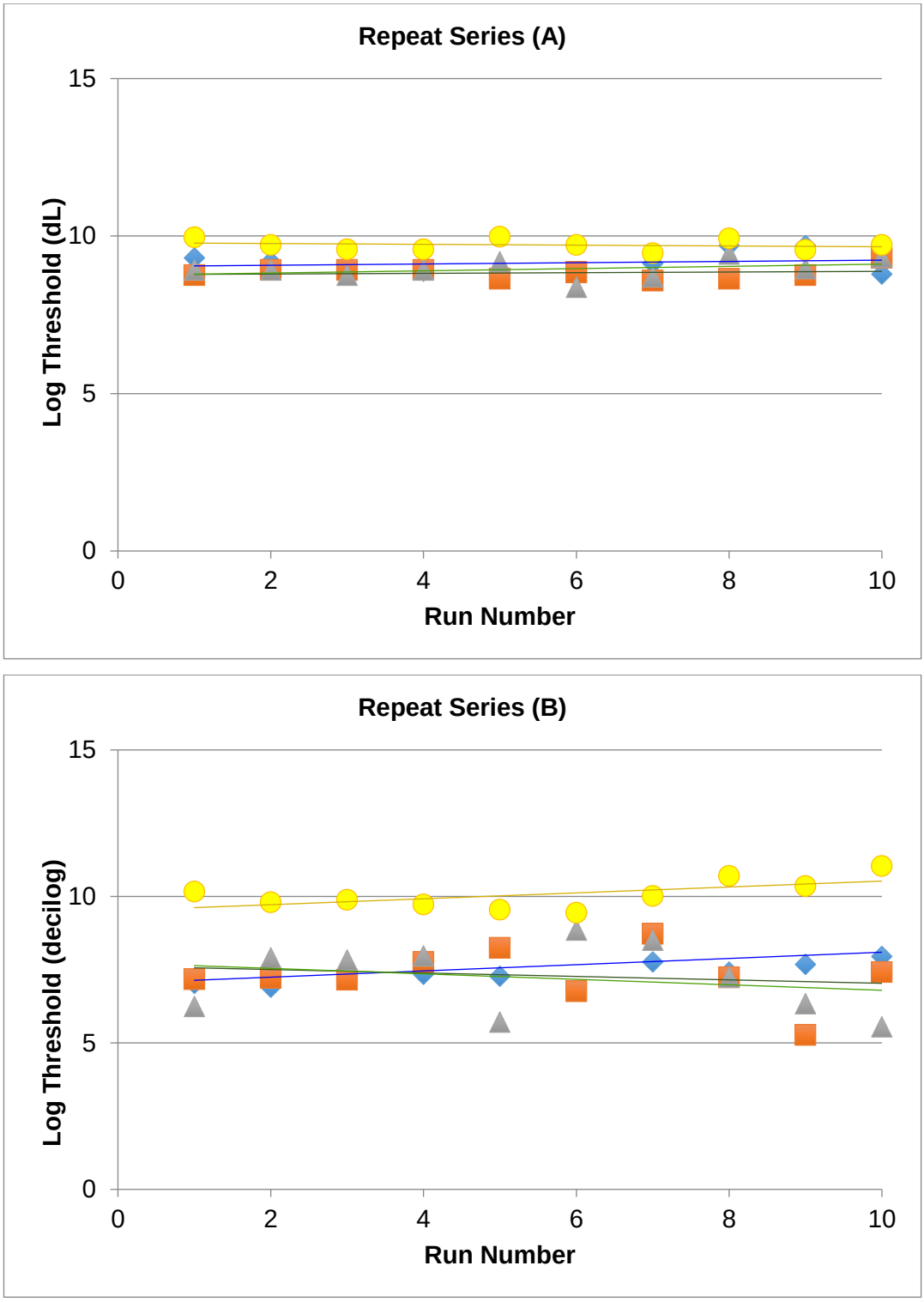


Figure 3.8. Trend analysis over 10 runs of the thresholds for each field location averaged for four control observers. **A:** Data for the criterion run length of 8 (T8 condition). **B:** Data for the criterion run length of 12 (T12 condition). None of the slopes differed significantly from zero ($p<0.05$).

The four control observers completed 10 runs over a four-month period to determine stability of the estimates. The average threshold values for the T8 and T12 are plotted in Figures. 3.8 A and B. For the T8 condition, all four slopes are less than 0.5 dL over the 10 runs, with small differences in the average sensitivity across observers and an average σ of 1.66 dL.

For the T12 condition, the data are generally more variable, with an average σ of 1.75 dL and somewhat lower threshold values, but there are again no significant trends for improvement with practice.

The test durations were significantly longer for T12 (mean 147 ± 40 s) compared to T8 (Mean 76 $\pm$ 15 s) p=<0.0001, indicating a clear advantage of the T8 condition in terms of testing efficiency. The total number of responses in the T8 condition averaged 42, implying that the stable threshold values were obtained in just over 10 responses per visual field location. The corresponding number for the T12 condition was 76. Thus, both in terms of time taken and of variability of the threshold estimate, there is a clear advantage for the T8 condition.

3.4. Commentary on the pre-clinical development of the A4FTp

As part of the pre-clinical development of the A4FTp we evaluated the psychometric properties of the test in 20 short-term and 4 long-term participants. These datasets both show that there was a significant increase in the time taken to reach the 12-point (T12) staircase criterion relative to the 8-point (T8) criterion. The average time for T8 was a little over one minute per eye for estimation of the four thresholds, while that for T12 was well over two minutes. There was also a significant learning effect on completion time for T12 (ranging from 166 s for the first session to 126 s for the fourth session) but not for T8. Thus the 8-point termination criterion has a clear advantage in terms of both efficiency of the A4FTp staircase method in achieving the threshold estimates and stability of its performance, with the absence of any learning effect for T8. The question is whether this speed advantage was obtained at the cost of a reduction in the quality of the resulting threshold estimates.

There are two relevant measures of the quality of the estimates, stability over time and variation across individuals. Within-test variability is caused by several aspects, neural noise, decision criteria, and thresholding strategy (Johnson et al

1988; Gonzalez de la Rosa et al 2006). Between-test variability has been attributed to ocular and neural sensitivity fluctuations (Spry et al 2002). The variability that occurs during test and retest is quantified by the degree of scatter between measurements taken at different test sessions (Chauhan et al 1999; Spry et al 2000; Spry et al 2002). Considering firstly the stability over time, we see that there was no significant drift in the estimated values for either criterion, for either the short- or long-term test series (even though the duration for T12 did significantly decrease over time). Thus, the A4FTp staircase was equally stable under all test conditions. However, there was a big difference between the criteria in the variation between individuals, which had the remarkably low standard deviation of 0.52 dL for T8 compared with 1.32 dL for T12. These values may be compared with the standard deviation values of ~1.7 dL reported for the population variation of two alternative forced choice flicker threshold estimates in previous studies (Tyler et al 1991). Thus, while the standard deviation for the 12-point criterion is close to the range recorded in previous studies, the 8-point criterion markedly reduces the standard deviation to about one third of this level implying that the physiological strain of the longer runs overcame the statistical advantage of the increased number of samples.

Since variance is defined as the square of the standard deviation, this result further implies that about 90% of the variance in the 12-point criterion staircase, was due to methodological variations, with no more than 10% of the variance attributable to inherent population variability. Moreover, there is no significant gain in reliability for the extra time spent to reach the 12-point criterion, and, though slightly lower on average, the threshold values themselves are not significantly different from those for the 8-point criterion. Thus, we have to conclude from the 20-participant results that, of the two approaches evaluated, the 8-point criterion staircase is a more effective approach to flicker threshold measurement, with no evident disadvantages and a clear time advantage over the 12-point criterion staircase.

Learning and fatigue effects are an important issue in many psychophysical tests and consist of an improvement or degradation, in performance respectively, as a function of the duration of the test. The learning effect occurs as the patient becomes increasingly familiar with the requirements of the perimetric task and manifests as an improvement in sensitivity and a decrease in measurement variability over time. This phenomenon tends to increase the false-positive rates

for inexperienced examinees, influencing the specificity, and can therefore be detrimental to the implementation of any technology used as a screening device.

Previous findings suggest that learning effects for Frequency Doubling Technology (FDT) are mild, regardless of the version of the device used or the screening strategy (Lester et al 2000; Khong et al 2001; Spry et al 2001; Fujimoto et al 2002; Horani et al 2002; Joson et al 2002; Matsuo et al 2002; Brush et al 2004; Hong et al 2007; Centofanti et al 2008). Learning effects are well known for SAP, in that subsequent examinations give increased absolute mean sensitivity over the initial session (Werner et al 1990; Heijl et al 1996; Schimiti et al 2002). Such learning effects were also identified in SWAP (Wild et al 1996; Rossetti et al 2006), and flicker perimetry (Bernardi et al 2007). It is impressive, therefore, that the A4FTp shows no learning effects in the present control sample, even up to 10 test repeats.

Testing times of the A4FTp (mean 76 ± 15s for T8 per eye) were faster than other glaucoma screening tests that use a threshold strategy, which range from 2.5 to 9 minutes per eye (Burr et al 2007). Most perimeters used for screening resort to a supra-threshold algorithm that sacrifices sensitivity for speed (Burr et al 2007).

In conclusion, this initial study showed that the T8 version of the A4FTp flicker sensitivity test has a shorter duration than equivalent threshold perimeters, did not show any statistically significant learning effect over multiple repetitions.

3.5. Diagnostic accuracy of the A4FTp for the detection of COAG

Purpose

The diagnostic accuracy of the A4FTp flicker test was determined using a case-control design were patients with an established diagnosis of COAG were compared with those of a control group. The A4FTp Test was also compared with two other tests for glaucoma case finding: The Frequency Doubling Technology (FDT) perimeter and the iVue Spectral Domain Optical Coherence Tomography (SD-OCT). User acceptability data was collected for all screening tests. The study was designed and reported, in accordance with the Standards for Reporting of Diagnostic Accuracy guidelines (STARD) (Bossuyt et al 2003).

Methods

Forty consecutive adults with a clinical diagnosis of COAG where recruited using the university eye clinic request and also via an advertisement in the International Glaucoma Association newsletter. The control group consisted of consecutive non-glaucomatous adults that were recruited from local optometry practices and the university eye clinic. Figure 3.9 shows the flow of patients through the study.

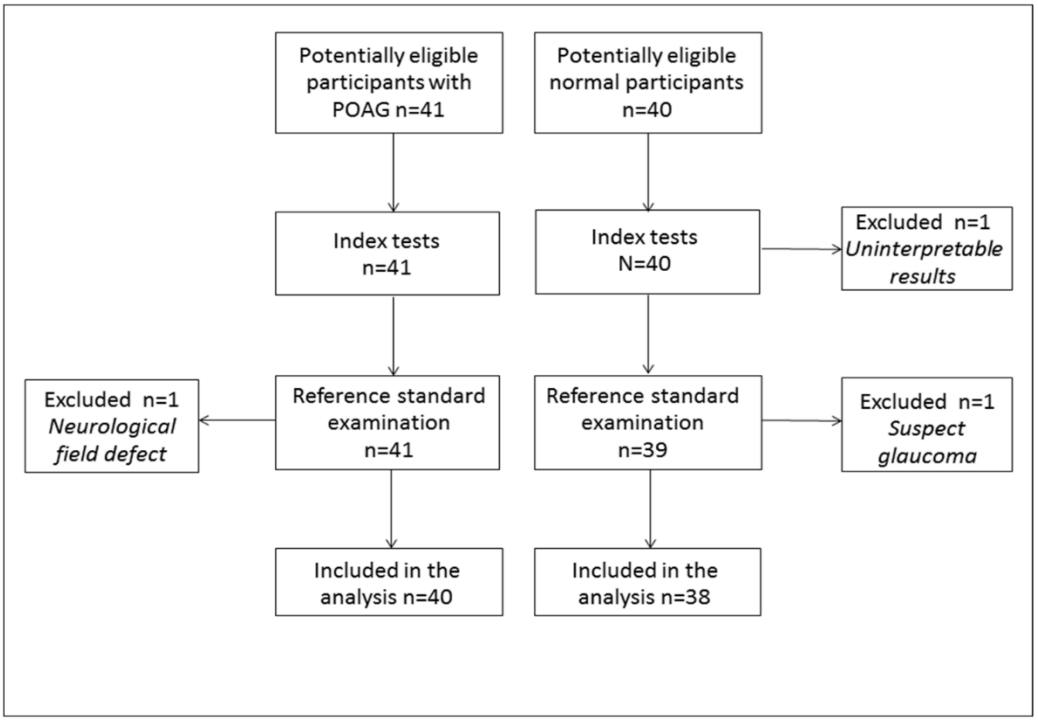


Figure 3.9. Study Flow Diagram

Reference standard

All participants underwent a reference standard ophthalmic examination on the same day as the index tests by an experienced glaucoma-specialist optometrist AJ (with training and accreditation within glaucoma clinics in the UK Hospital Eye Service), masked to the index tests results. The reference examination included Log MAR best correct visual acuity, refraction, IOP using a Goldmann Applanation Tonometer (GAT), slit-lamp biomicroscopy, grading of the crystalline lens using LOCS II, van Herick assessment of limbal anterior chamber depth (van Herick et al 1969), potentially occludable angles examined using gonioscopy, and dilated fundus examination with detailed disc assessment using indirect ophthalmoscopy. Visual fields were assessed using a Carl Zeiss Humphrey Field Analyzer (HFA) with the Swedish Interactive Thresholding Algorithm 24-2 standard pattern (SITA Standard). HFA field-testing was repeated for false positives >15%, false negatives or fixation losses >33%.

Inclusion and exclusion criteria

Adults aged 40 and above were included in the study, control participants included in the study where free from ocular disease with normal appearance of the optic disc, normal fundus, IOP ≤21 mmHg, and full visual fields. COAG diagnosis was based on the following criteria: open anterior chamber angles; presence of glaucomatous optic neuropathy (localised absence of neuroretinal rim, cup-to-disc ratio 0.7, or interocular asymmetry in vertical cup-to-disc ratio 0.2 in similar sized discs); the presence of a concordant glaucomatous field defect using the 24-2 SITA algorithm on the HFA. Glaucomatous visual field loss was classified using Hodapp-Parrish Anderson criteria as early, moderate or advanced (Hodapp et al 1993).

Participants were excluded if they had a history of angle closure, significant diabetic retinopathy, retinal vascular occlusions, peripheral retinal abnormalities, optic atrophy, clinically significant cataract [Lens Opacity Classification System III (LOCIII)] (Chylack et al 1993), or a neurological field defect.

Index tests

All participants underwent testing with all three index tests, A4FTp, FDT (using the C20-5 programme), and iVue OCT (RNFL and total retinal thickness), performed in a random order by BF who was unaware of the participants' ocular status. Thresholds of abnormality for the index tests were based on cut-offs commonly reported in the literature and were predefined before data analysis.

Statistical Analysis

Index and reference tests were performed on both eyes. However, for the purpose of the analysis, data from one eye was used. In the case of participants with glaucoma, this was the eye with the greater visual field loss (given that the design of the A4FTp was optimised to detect more established field loss). For consistency, the right eye was selected for the controls; if the right eye was not eligible, the left eye was used.

T-tests were used to compare differences between groups to evaluate diagnostic test accuracy, Receiver Operator Characteristics (ROC) curves were plotted. Differences in the area under the ROC curve (AUROC) for each test parameter at the 95% confidence interval (CI) were compared statistically using the DeLong method (DeLong et al 1988). Sensitivity, specificity, positive and negative likelihood ratios were also calculated.

Results

Of the 81 invited participants in the study 3 where excluded from the analysis; two control participants one with uninterpretable test results on all visual function tests and another was identified as a COAG suspect, one COAG participant had bilateral neurological visual field defects. Table 3.2 provides a summary of the demographic and clinical data for the controls and COAG participants. Statistical differences between groups were found for age, IOP and HFA SITA 24-2 threshold MD.

Table 3.2. Demographical and summary clinical data				
	Overall	Control	COAG	p
No. participants	78	38	40	
Age (Mean ± SD) (years)	66.8 ± 11.2	61.6 ± 10.6	71.9 ± 9.4	<0.001
Female No. (%)	46 (59%)	22 (57.9%)	24 (60%)	0.85
Ethnicity				
Caucasian No. (%)	65 (83.3%)	28 (73.7%)	37 (92.5%)	
Asian Indian No. (%)	12 (15.4%)	10 (26.3%)	2 (5%)	
African origin No. (%)	1 (1.3%)	0 (0%)	1 (2.5%)	
Visual acuity (Log) (Mean ± SD)	0.07 ± 0.15	0.04 ± 0.17	0.09 ± 0.12	0.092
IOP (mmHg) (Mean ± SD)	16.4 ± 4.42	17.5 ± 2.5	15.3 ± 5.5	0.024
Refractive error (DS) (Mean ± SD)	0.50 ± 3.20	0.26 ± 3.59	0.72 ± 2.81	0.53
Refractive error (DC) (Mean ± SD)	0.87 ± 0.84	0.72 ± 0.71	1.01 ± 0.93	0.13
HFA SAP SITA 24-2 threshold MD (dB) (Mean ± SD)	5.75 ± 7.41	0.71 ± 1.55	10.53 ± 7.61	<0.001

DC, Dioptric Cylinder; DS, Dioptric Sphere; MD, Mean Deviation; PSD, Pattern Standard Deviation; SAP, Standard automated perimetry: SD, Standard deviation.

Figure 3.10 shows a histogram of the mean log flicker thresholds for the control and COAG subgroups with their 95% confidence intervals. As disease severity increased, there was a rise of the mean log threshold for flicker detection. There was overlap in the distribution for control participants and those with early COAG. Consequently, the test failed to identify almost half (n=6, 46%) of patients with early COAG, while correctly identifying 93% and 100% of moderate and severe COAG respectively. Lowering the log threshold could increase sensitivity at a cost of lowering specificity as shown on Figure 3.11.

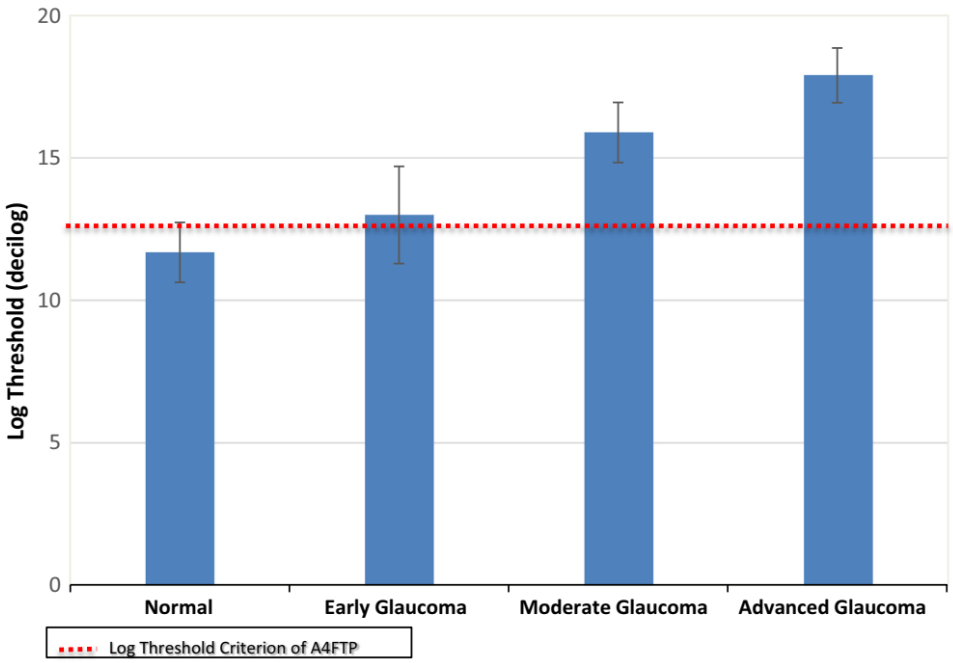


Figure 3.10. Histogram of the mean log thresholds of control and glaucoma subgroups. Glaucomatous visual field loss was classified using Hodapp-Parrish Anderson criteria (Hodapp et al 1993).

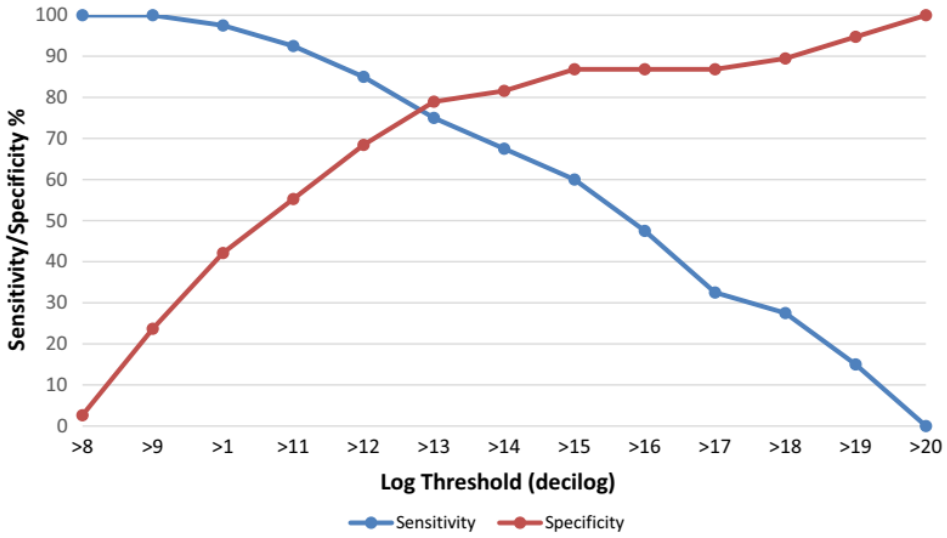


Figure 3.11. Sensitivity and specificity plot for A4FTp.

Table 3.3. Pairwise comparison of ROC curves of the index tests			
Test comparisons	Difference between the areas	95% Confidence interval	P
A4FTp vs FDT p < 1% any point missed	0.08	-0.03 to 0.18	0.15
A4FTp vs FDT p < 5% any point missed	0.09	-0.02 to 0.02	0.12
A4FTp vs SD-OCT p < 1%	0.07	-0.04 to 0.18	0.18
A4FTp vs SD-OCT p < 5%	0.01	-0.11 to 0.12	0.91

Table 3.3 shows a comparison of the AUROC curves between the three index tests. No statistically significant differences where identified between the A4FTp and FDT p < 1% level or FDT p < 5% level. A comparison between the A4FTp and the SD-OCT RNFL (any quadrant) at the p < 1% level or p < 5% level also failed to identify any statistically significant differences.

The mean AUROC for the three tests were; A4FTp (0.824, 95% confidence interval (0.726-0.921)), SD-OCT (any RNFL parameter p<1% level) (0.898 (0.830-0.966)) and FDT (one or more locations missed at p<5% level) (0.911 (0.824-0.963)).

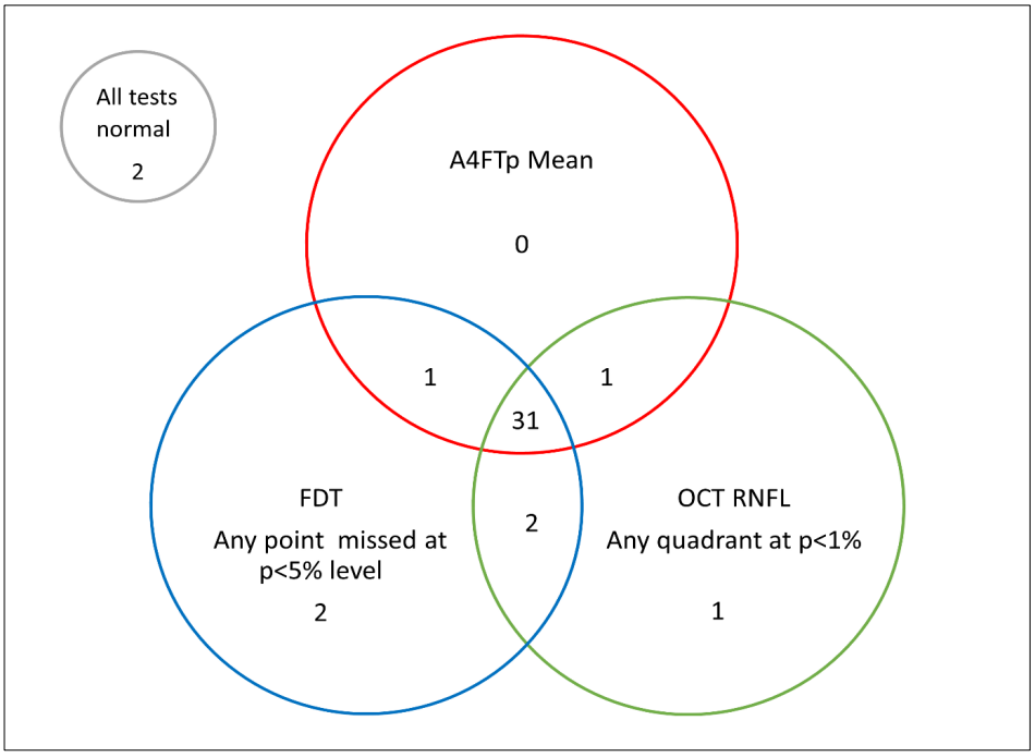


Figure 3.12. Venn diagram of best performing parameter from the index tests in identifying the COAG cases alone or combined with the other tests.

Figure 3.12 shows a Venn diagram for the best performing criteria of each index test for detecting COAG participants. The numbers within the circles represent the number of participants identified by each test. The A4FTp detected slightly fewer COAG cases (n=33, 83%) than the FDT (n=36, 90%) or SD-OCT (n=35, 88%). All three index tests failed to defect two cases (5%). The diagram reveals that combining a structural test (SD-OCT) with a functional test (FDT or A4FTp) increases the likelihood of detecting the disease.

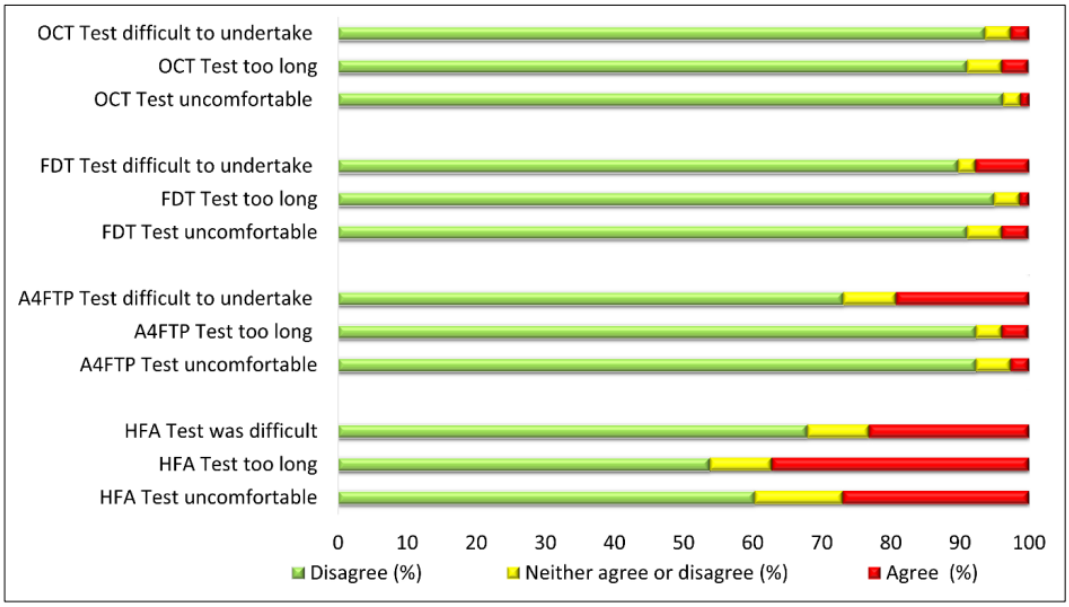


Figure 3.13. Likert responses from user acceptability survey

Figure 3.13 shows the Likert scores from the tests from the acceptability survey. The A4FTp, FDT and SD-OCT had a similar proportion of participants (>90%) rating the tests as not uncomfortable or not too long. Of the four tests the HFA was flagged as uncomfortable, too long and difficult to perform by comparison. Although the A4FTp was considered comfortable and relative short some participants found the A4FTp difficult to perform in its current configuration.

Strengths and limitations of the study

Determination of diagnostic accuracy is a key step in the development of a new screening test. Diagnostic accuracy is the ability of a test to differentiate between patients who have the condition of interest (target condition) and those who do not. The accuracy of the test is evaluated by comparing the results of the test with an established diagnostic reference standard on the same series of participants. Although the case-control design used in the current study is likely to overestimate diagnostic accuracy, it is a convenient first stage in the evaluation of a new test. The design also allows for an assessment of the comparative accuracy of the new test with more established screening tests for COAG.

The A4FTp has a number of advantages, including its ease of administration and interpretation, a relatively short testing time and robustness to the effects of fixation losses, media opacity or refractive error. Since flicker sensitivity is less affected by age than standard automated perimetry (Tyler 1989; Lachenmayr et al 1994), we did not attempt to age match the cases and controls. Consequently, there was a statistically significant difference in the mean age of the two groups. Although we cannot exclude the possibility that this baseline age difference may have impacted on the observed difference in flicker sensitivity between groups, the same confounder would have applied to the FDT used for the comparative analysis.

Another potential limitation of the A4FTp was its failure to detect 6 out of the 12 patients with early glaucoma. The initial design of the A4FTp tested a small number of locations to detect those with more advanced visual field loss, on the basis that those at higher risk of significant visual disability in their lifetime would generally present with greater field damage at presentation (Saunders et al., 2014). A test strategy could potentially be developed to improve the ability of the A4FTp to detect early glaucoma by using smaller stimuli and further test locations.

Conclusion

Based on a comparison of the AUROC curve, the overall performance of the A4FTp was similar to the FDT (C20-5 algorithm) and the SD-OCT (RNFL thickness outside normal limits). The best performing criterion for the A4FTp was the mean threshold of all four stimulus locations. The optimal threshold criterion for the A4FTp was based on an equal weighting for sensitivity and specificity. Using this criterion, the test identified 33 out of the 40 glaucoma cases in our sample (83%).

Test accuracy for all index tests was equivalent for the detection of COAG. Time taken to complete the A4FTp was relatively short and initial results are promising. User acceptability of the A4FTp was positive, with similar acceptability questionnaire scores to the FDT and SD-OCT, in terms of comfort and participant's opinion on the duration of the test. Although some participants found the A4FTp difficult to perform, this was generally related to issues relating to identifying the appropriate key on the keypad corresponding to the location of the presented stimulus. However, further refinement and optimisation of the A44FTp in the future could include the use of touch screen technology, which would remove the need for the keypad.

3.6. Role in the study

Professor Christopher Tyler wrote the original code in MATLAB for the stimulus algorithm. Under Professor Tyler's supervision, I coded the version of the algorithm used in the A4FTp, including stimulus size, location, input configuration, working database and output files. I then designed the validation study, recruited participants, collected data, and analysed the data in consultation with Professors Tyler and Lawrenson. The preliminary findings were reported via poster at BCOVS 2015 before the subsequent diagnostic accuracy study was performed. The diagnostic accuracy study was conducted jointly with another PhD student (Anish Jindal) under Professor Lawrenson's supervision. The findings were reported at the European Association for Vision and Eye Research (EVER) meeting in 2018. My role in this part of the study included data collection, analysis and I jointly drafted the submitted manuscript to Ophthalmic and Physiological Optics with Anish Jindal.

My percentage contribution to this study: 70%

3.7. References

Anderson, D.R., Patella, V.M., 1999. Automated static perimetry, 2nd ed. ed. Mosby, St. Louis.

Bernardi, L., Costa, V.P., Shiroma, L.O., 2007. Flicker perimetry in healthy subjects: influence of age and gender, learning effect and short-term fluctuation. Arquivos Brasileiros de Oftalmologia 70, 91–99.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Moher, D., Rennie, D., de Vet, H.C.W., Lijmer, J.G., Standards for Reporting of Diagnostic Accuracy, 2003. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Annals of Internal Medicine 138, W1-12.

Bourne, R.R.A., Stevens, G.A., White, R.A., Smith, J.L., Flaxman, S.R., Price, H., Jonas, J.B., Keeffe, J., Leasher, J., Naidoo, K., Pesudovs, K., Resnikoff, S., Taylor, H.R., 2013. Causes of vision loss worldwide, 1990–2010: a systematic analysis. The Lancet Global Health 1, e339–e349.

Brush, M.B., Chen, P.P., 2004. Learning effect among perimetric novices with screening C-20-1 frequency doubling technology perimetry. American Journal of Ophthalmology 137, 551–552.

Burr, J.M., Mowatt, G., Hernández, R., Siddiqui, M. a. R., Cook, J., Lourenco, T., Ramsay, C., Vale, L., Fraser, C., Azuara-Blanco, A., Deeks, J., Cairns, J., Wormald, R., McPherson, S., Rabindranath, K., Grant, A., 2007. The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation. Health Technology Assessment 11, iii–iv, ix–x, 1–190.

Centofanti, M., Fogagnolo, P., Oddone, F., Orzalesi, N., Vetrugno, M., Manni, G., Rossetti, L., 2008. Learning effect of humphrey matrix frequency doubling technology perimetry in patients with ocular hypertension. Journal of Glaucoma 17, 436–441.

Chauhan, B.C., Johnson, C.A., 1999. Test-retest variability of frequency-doubling perimetry and conventional perimetry in glaucoma patients and normal subjects. Investigative Ophthalmology and Visual Science 40, 648–656.

Chylack, L.T., Wolfe, J.K., Singer, D.M., Leske, M.C., Bullimore, M.A., Bailey, I.L., Friend, J., McCarthy, D., Wu, S.Y., 1993. The Lens Opacities Classification System III. The Longitudinal Study of Cataract Study Group. Archives of Ophthalmology 111, 831–836.

Contestabile, M.T., Perdicchi, A., Amodeo, S., Recupero, V., Recupero, S.M., 2007. The influence of learning effect on frequency doubling technology perimetry (Matrix). Journal of Glaucoma 16, 297–301.

De la Rosa, M.G., Reyes, J.A.A., Sierra, M.A.G., 1990. Rapid assessment of the visual field in glaucoma using an analysis based on multiple correlations. Graefe's Archive for Clinical and Experimental Ophthalmology 228, 387–391.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Fujimoto, N., Minowa, K., Miyauchi, O., Hanawa, T., Adachi-Usami, E., 2002. Learning effect for frequency doubling perimetry in patients with glaucoma. American Journal of Ophthalmology 133, 269–270.

Gonzalez de la Rosa, M., Gonzalez-Hernandez, M., Garcia Feijoo, J., Morales, J., Azuara-Blanco, A., 2006. Diagnostic accuracy and reproducibility of tendency oriented perimetry in glaucoma. European Journal of Ophthalmology 16, 259–267.

Grigsby, S.S., Vingrys, A.J., Benes, S.C., King-Smith, P.E., 1991. Correlation of chromatic, spatial, and temporal sensitivity in optic nerve disease. Investigative Ophthalmology and Visual Science 32, 3252–3262.

Heijl, A., Bengtsson, B., 1996. The effect of perimetric experience in patients with glaucoma. Archives of Ophthalmology 114, 19–22.

Henson, D.B., Chauhan, B.C., Hobley, A., 1988. Screening for glaucomatous visual field defects: the relationship between sensitivity, specificity and the number of test locations. Ophthalmic and Physiological Optics 8, 123–127.

Hernández, R., Rabindranath, K., Fraser, C., Vale, L., Blanco, A.A., Burr, J.M., OAG Screening Project Group, 2008. Screening for open angle glaucoma: systematic review of cost-effectiveness studies. Journal of Glaucoma 17, 159–168.

Hodapp, E., Parrish, R.K., Anderson, D.R., 1993. Clinical decisions in glaucoma. Mosby, St. Louis, Mo.

Hong, S., Na, K., Kim, C.Y., Seong, G.J., 2007. Learning effect of Humphrey Matrix perimetry. Canadian Journal of Ophthalmology 42, 707–711.

Horani, A., Frenkel, S., Yahalom, C., Farber, M.D., Ticho, U., Blumenthal, E.Z., 2002. The learning effect in visual field testing of healthy subjects using frequency doubling technology. Journal of Glaucoma 11, 511–516.

Horn, F.K., Jonas, J.B., Korth, M., Jünemann, A., Gründler, A., 1997. The full-field flicker test in early diagnosis of chronic open-angle glaucoma. American Journal of Ophthalmology 123, 313–319.

Johnson, C.A., Adams, C.W., Lewis, R.A., 1988. Fatigue effects in automated perimetry. Applied Optics 27, 1030–1037.

Joson, P.J., Kamantigue, M.E.G., Chen, P.P., 2002. Learning effects among perimetric novices in frequency doubling technology perimetry. Ophthalmology 109, 757–760.

Keltner, J.L., Johnson, C.A., Cello, K.E., Edwards, M.A., Bandermann, S.E., Kass, M.A., Gordon, M.O., Ocular Hypertension Treatment Study Group, 2003. Classification of visual field abnormalities in the ocular hypertension treatment study. Archives of Ophthalmology 121, 643–650.

Kerrigan-Baumrind, L.A., Quigley, H.A., Pease, M.E., Kerrigan, D.F., Mitchell, R.S., 2000. Number of ganglion cells in glaucoma eyes compared with threshold visual field tests in the same persons. Investigative Ophthalmology and Visual Science 41, 741–748.

Khong, J.J., Dimitrov, P.N., Rait, J., McCarty, C.A., 2001. Can the specificity of the FDT for glaucoma be improved by confirming abnormal results? Journal of Glaucoma 10, 199–202.

Krakau, C.E., 1989. Visual field testing with reduced sets of test points. A computerized analysis. Documenta Ophthalmologica 73, 71–80.

Lachenmayr, B.J., Drance, S.M., 1992. The selective effects of elevated intraocular pressure on temporal resolution. German Journal of Ophthalmology 1, 26–31.

Lester, M., Capris, P., Pandolfo, A., Zingirian, M., Traverso, C.E., 2000. Learning effect, short-term fluctuation, and long-term fluctuation in frequency doubling technique. American Journal of Ophthalmology. 130, 160–164.

Matsuo, H., Tomita, G., Suzuki, Y., Araie, M., 2002. Learning effect and measurement variability in frequency-doubling technology perimetry in chronic open-angle glaucoma. Journal of Glaucoma 11, 467–473.

McKendrick, A.M., 2005. Recent developments in perimetry: test stimuli and procedures. Clinical and Experimental Optometry 88, 73–80.

Medeiros, F.A., Lisboa, R., Weinreb, R.N., Liebmann, J.M., Girkin, C., Zangwill, L.M., 2013. Retinal ganglion cell count estimates associated with early development of visual field defects in glaucoma. Ophthalmology 120, 736–744.

Mowatt, G., Burr, J.M., Cook, J.A., Siddiqui, M.A.R., Ramsay, C., Fraser, C., Azuara-Blanco, A., Deeks, J.J., OAG Screening Project, 2008. Screening tests for detecting open-angle glaucoma: systematic review and meta-analysis. Investigative Ophthalmology and Visual Science 49, 5373–5385.

Nicholas, S.P., Werner, E.B., 1980. Location of early glaucomatous visual field defects. Canadian Journal of Ophthalmology 15, 131–133.

Plant, G.T., Hess, R.F., 1987. Regional threshold contrast sensitivity within the central visual field in optic neuritis. Brain 110 (Pt 2), 489–515.

Rossetti, L., Fogagnolo, P., Miglior, S., Centofanti, M., Vetrugno, M., Orzalesi, N., 2006. Learning effect of short-wavelength automated perimetry in patients with ocular hypertension. Journal of Glaucoma 15, 399–404.

Saunders, L.J., Russell, R.A., Kirwan, J.F., McNaught, A.I., Crabb, D.P., 2014. Examining visual field loss in patients in glaucoma clinics during their predicted remaining lifetime. Invest. Ophthalmol. Vis. Sci. 55, 102–109.

Schimiti, R.B., Avelino, R.R., Kara-José, N., Costa, V.P., 2002. Full-threshold versus Swedish Interactive Threshold Algorithm (SITA) in normal individuals undergoing automated perimetry for the first time. Ophthalmology 109, 2084–2092; discussion 2092.

Spry, P.G., Henson, D.B., Sparrow, J.M., North, R.V., 2000. Quantitative comparison of static perimetric strategies in early glaucoma: test-retest variability. Journal of Glaucoma 9, 247–253.

Spry, P.G., Johnson, C.A., McKendrick, A.M., Turpin, A., 2001. Variability components of standard automated perimetry and frequency-doubling technology perimetry. Investigative Ophthalmology and Visual Science 42, 1404–1410.

Spry, P.G.D., Johnson, C.A., 2002. Identification of progressive glaucomatous visual field loss. Survey of Ophthalmology 47, 158–173.

Sugimoto, K., Schötzau, A., Bergamin, O., Zulauf, M., 1998. Optimizing distribution and number of test locations in perimetry. Graefe's Archive for Clinical and Experimental Ophthalmology 236, 103–108.

Tyler, C.W., 1981. Specific deficits of flicker sensitivity in glaucoma and ocular hypertension. Investigative Ophthalmology and Visual Science 20, 204–212.

Tyler, C.W., 1989. Two processes control variations in flicker sensitivity over the life span. Journal of the Optical Society of America A 6, 481–490.

Tyler, C.W., 1991. Analysis of normal flicker sensitivity and its variability in the visuogram test. Investigative Ophthalmology and Visual Science 32, 2552–2560.

van Herick, W., Shaffer, R.N., Schwartz, A., 1969. Estimation of Width of Angle of Anterior Chamber. American Journal of Ophthalmology 68, 626–629.

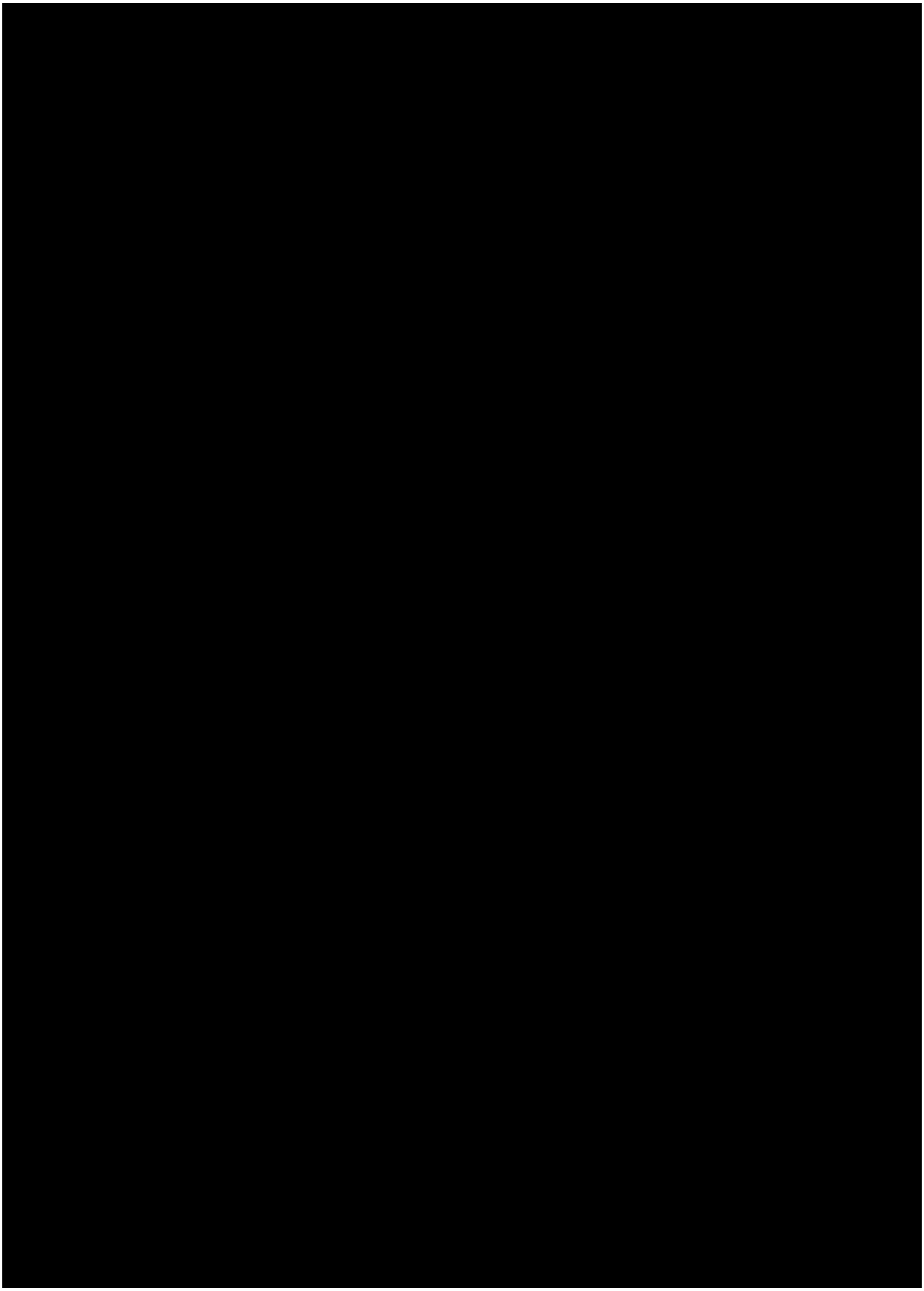
Wang, Y., Henson, D.B., 2013. Diagnostic Performance of Visual Field Test Using Subsets of the 24-2 Test Pattern for Early Glaucomatous Field Loss. Investigative Opthalmology & Visual Science 54, 756.

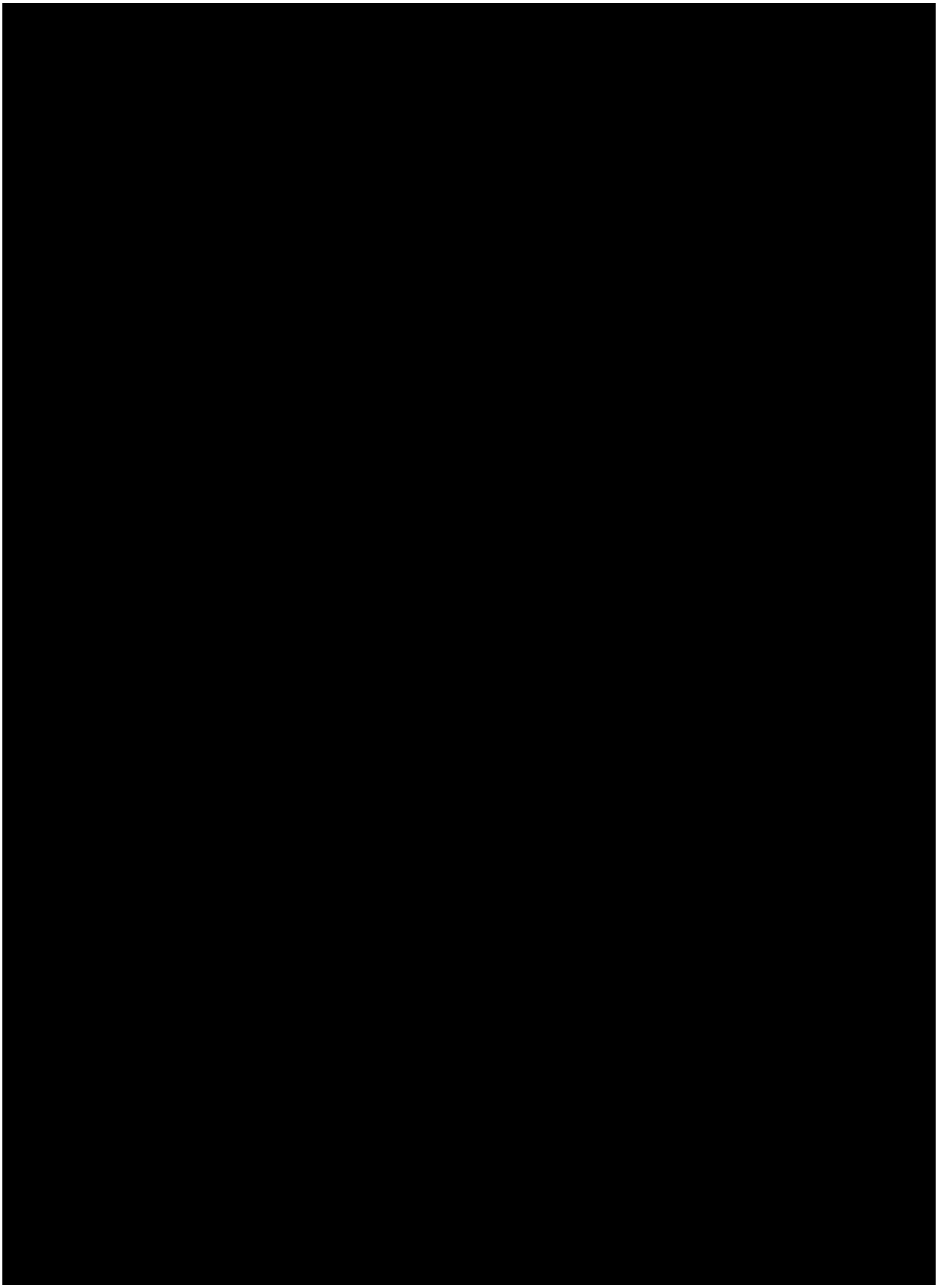
Werner, E.B., Krupin, T., Adelson, A., Feitl, M.E., 1990. Effect of patient experience on the results of automated perimetry in glaucoma suspect patients. Ophthalmology 97, 44–48.

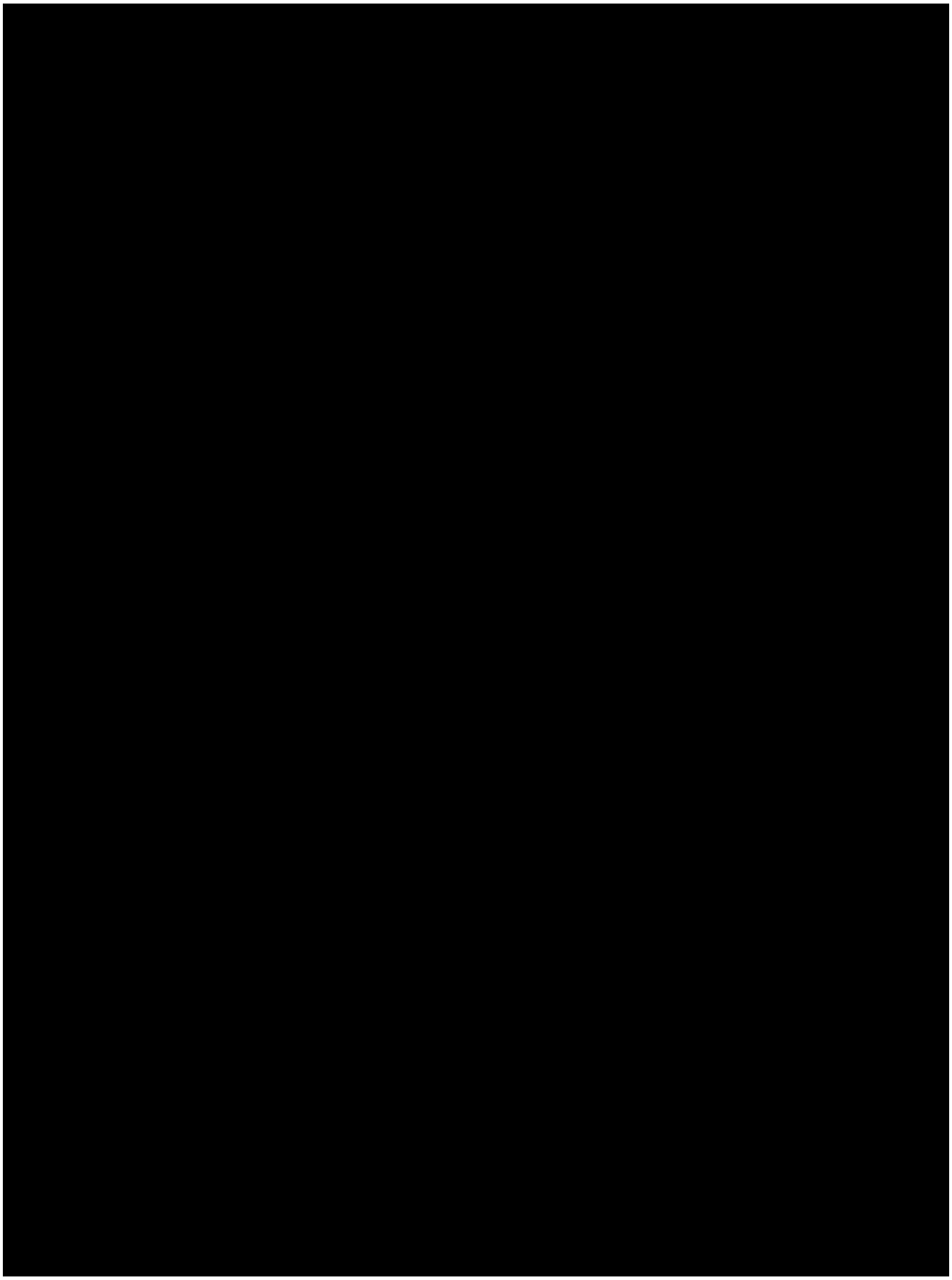
Wild, J.M., Moss, I.D., 1996. Baseline alterations in blue-on-yellow normal perimetric sensitivity. Graefe's Archive for Clinical and Experimental Ophthalmology 234, 141–149.

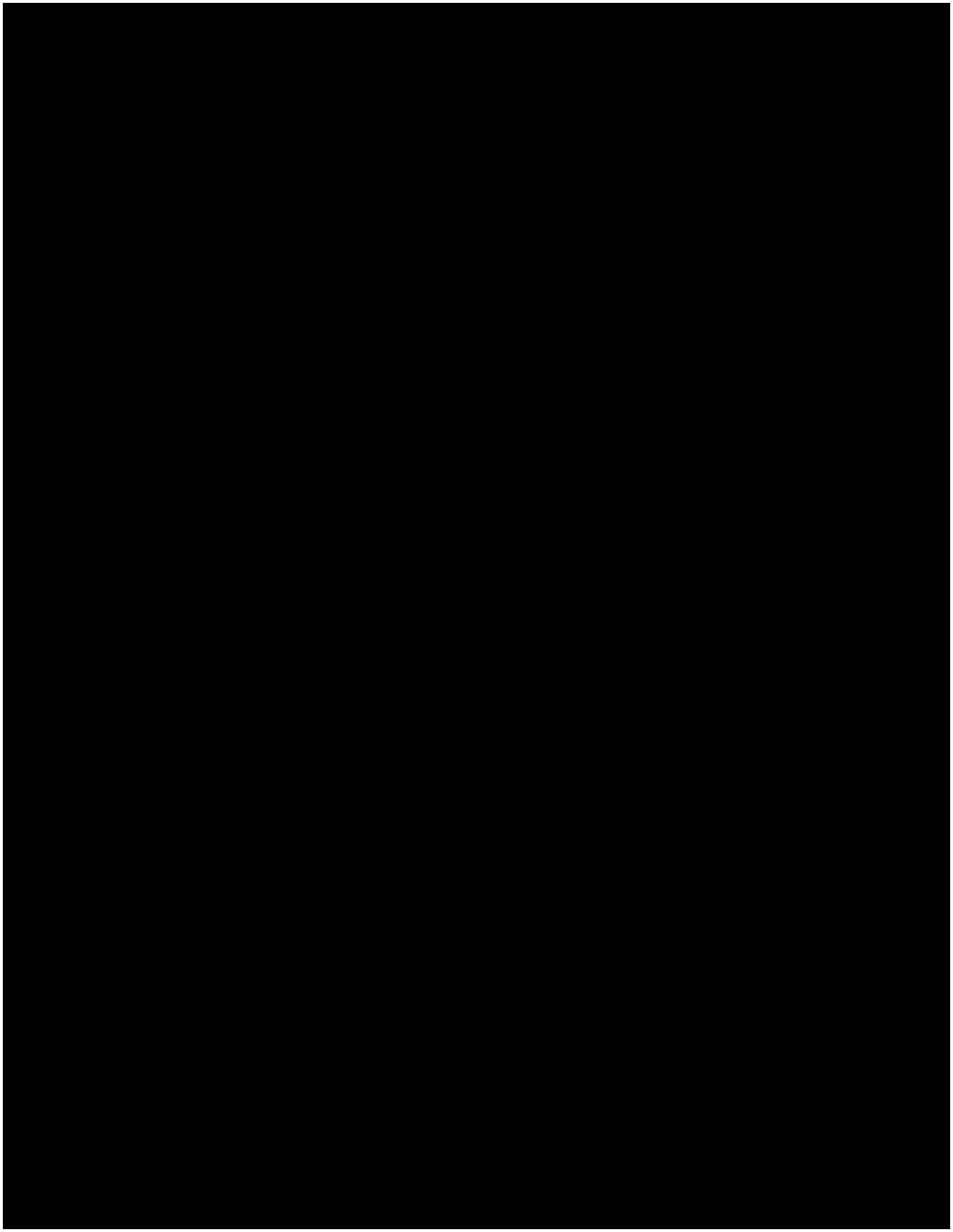
Wilson, J.M.G., Jungner, G., 1968. The Principles and Practice of Screening for Disease, Public Health Papers 34. WHO, Geneva.

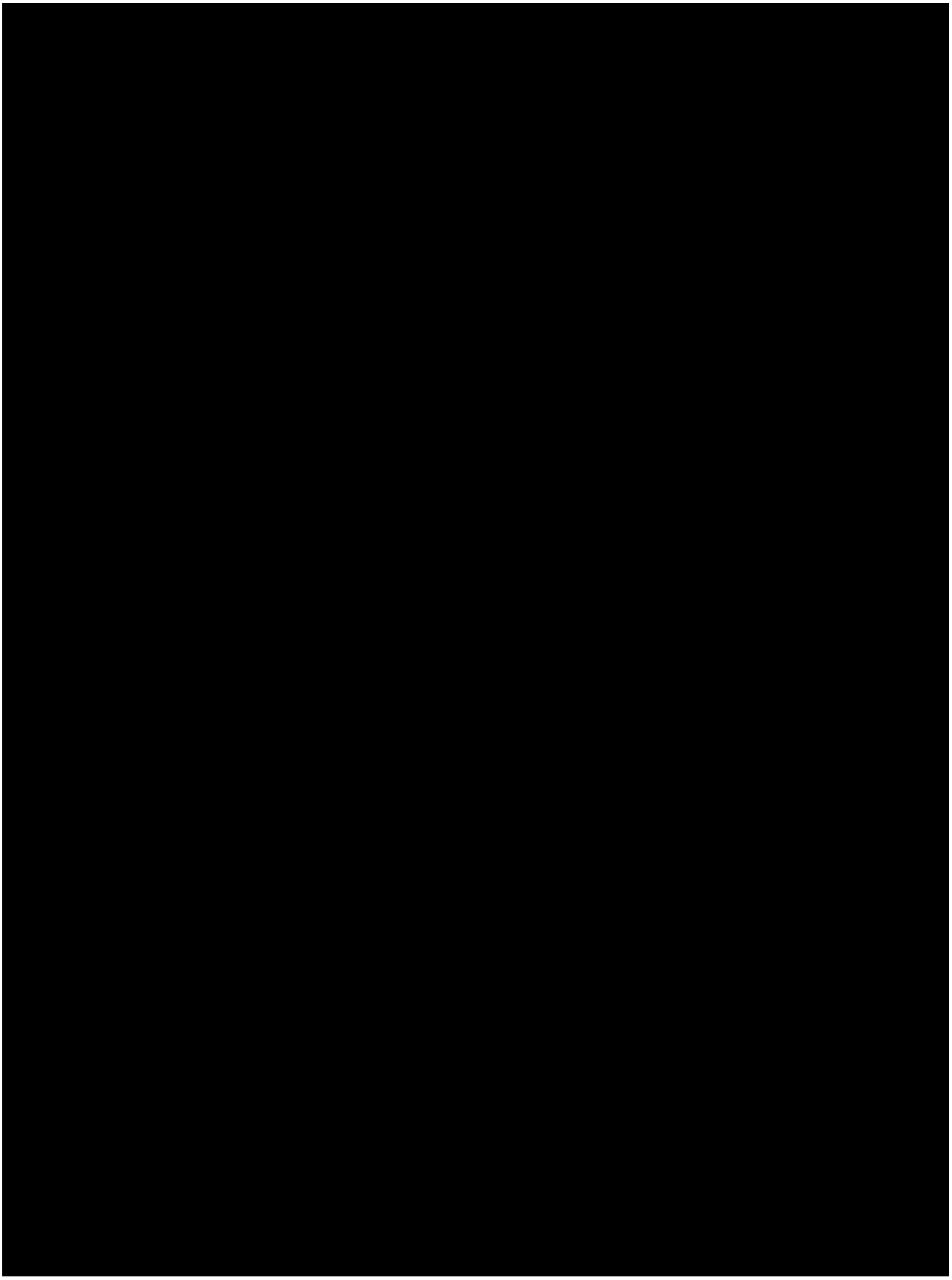
3.8. Published Article

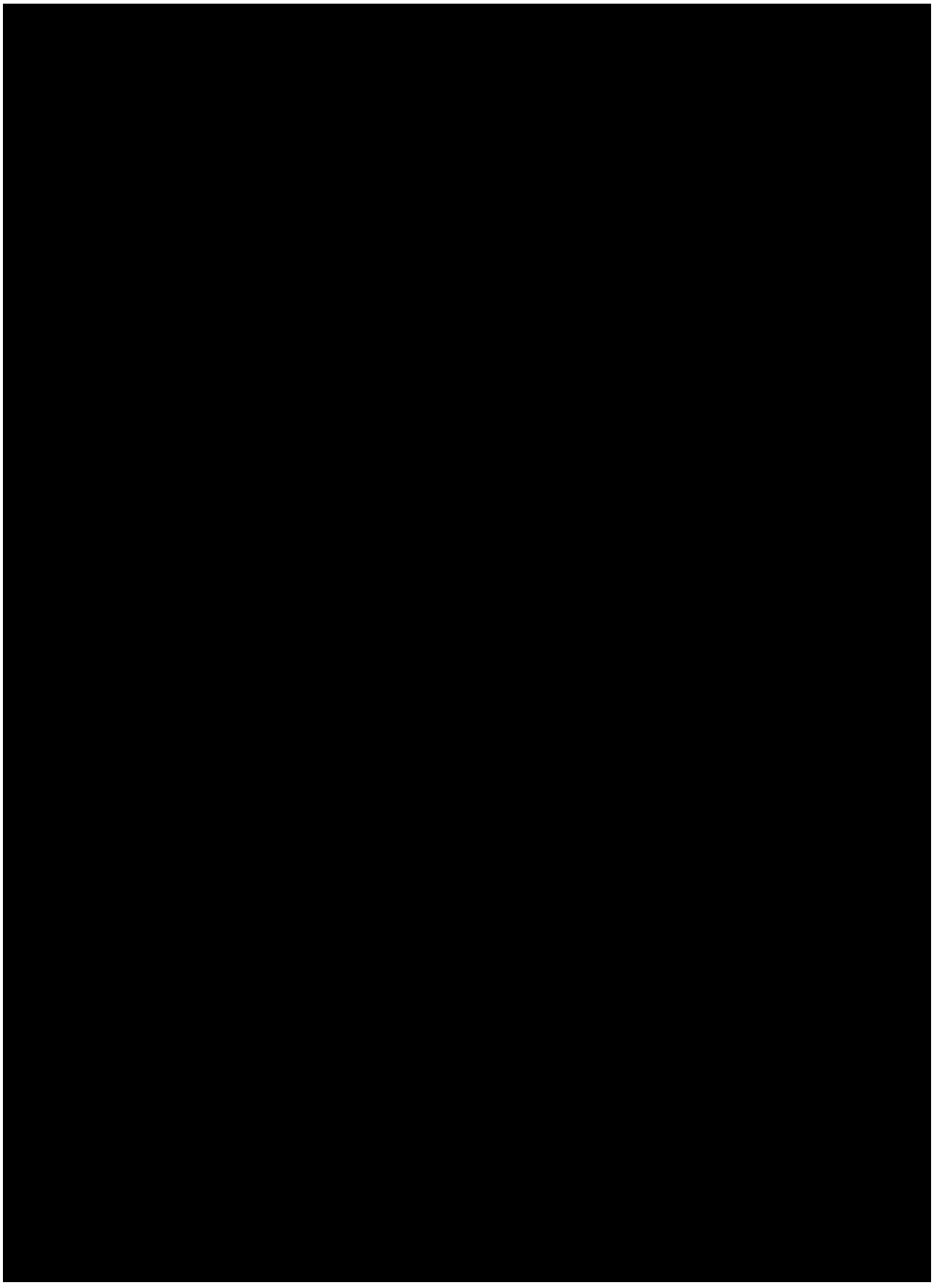


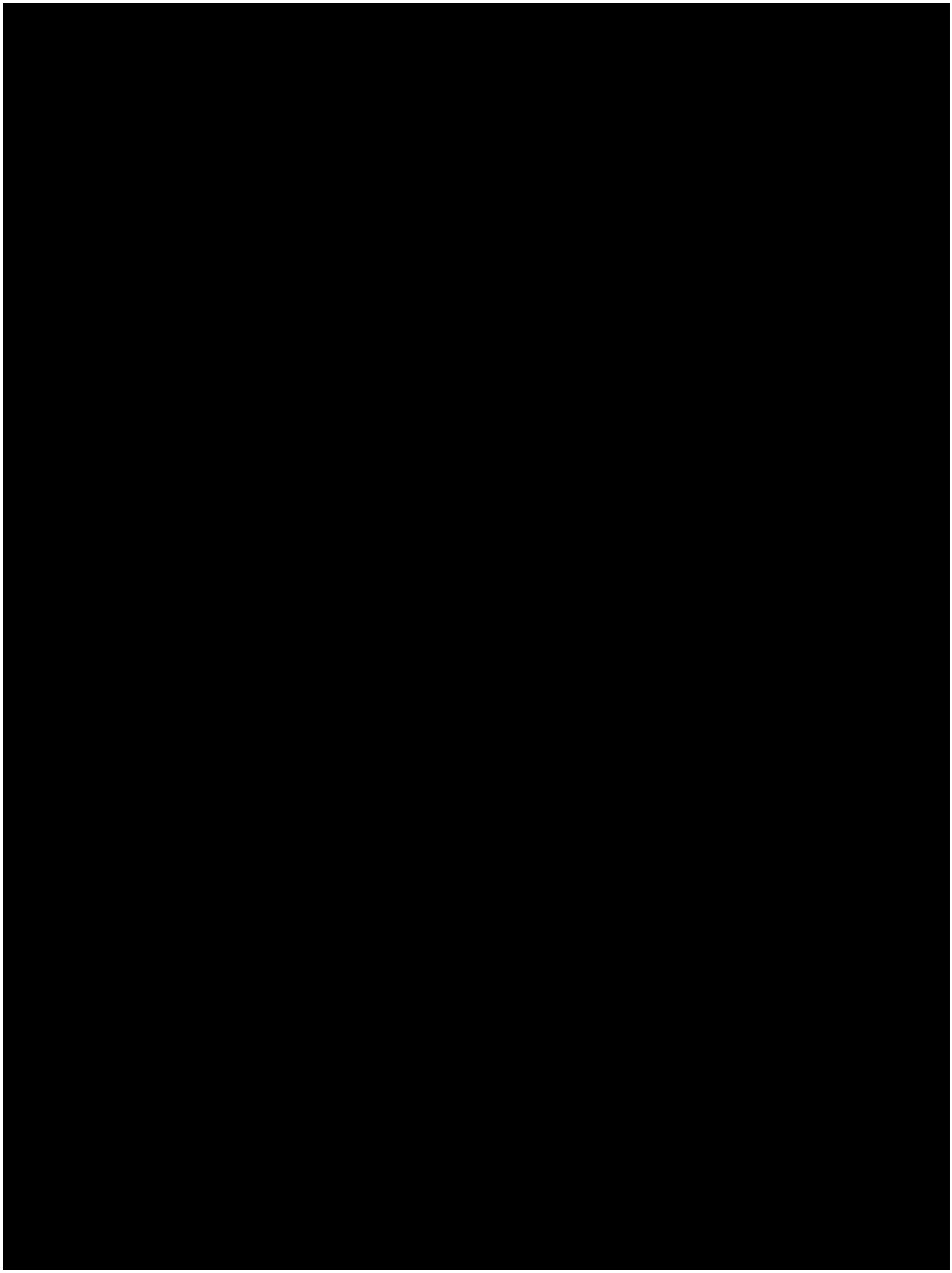


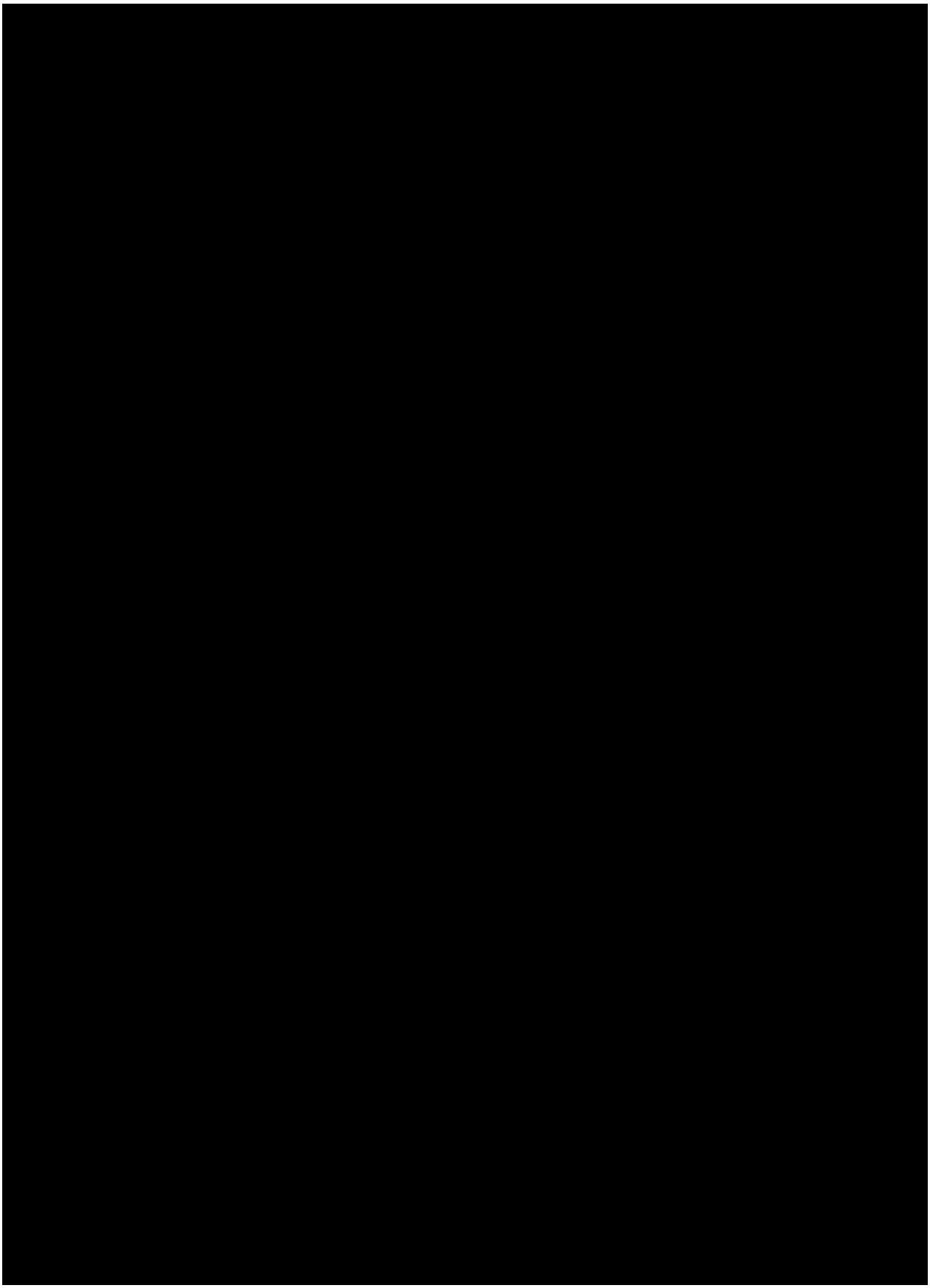


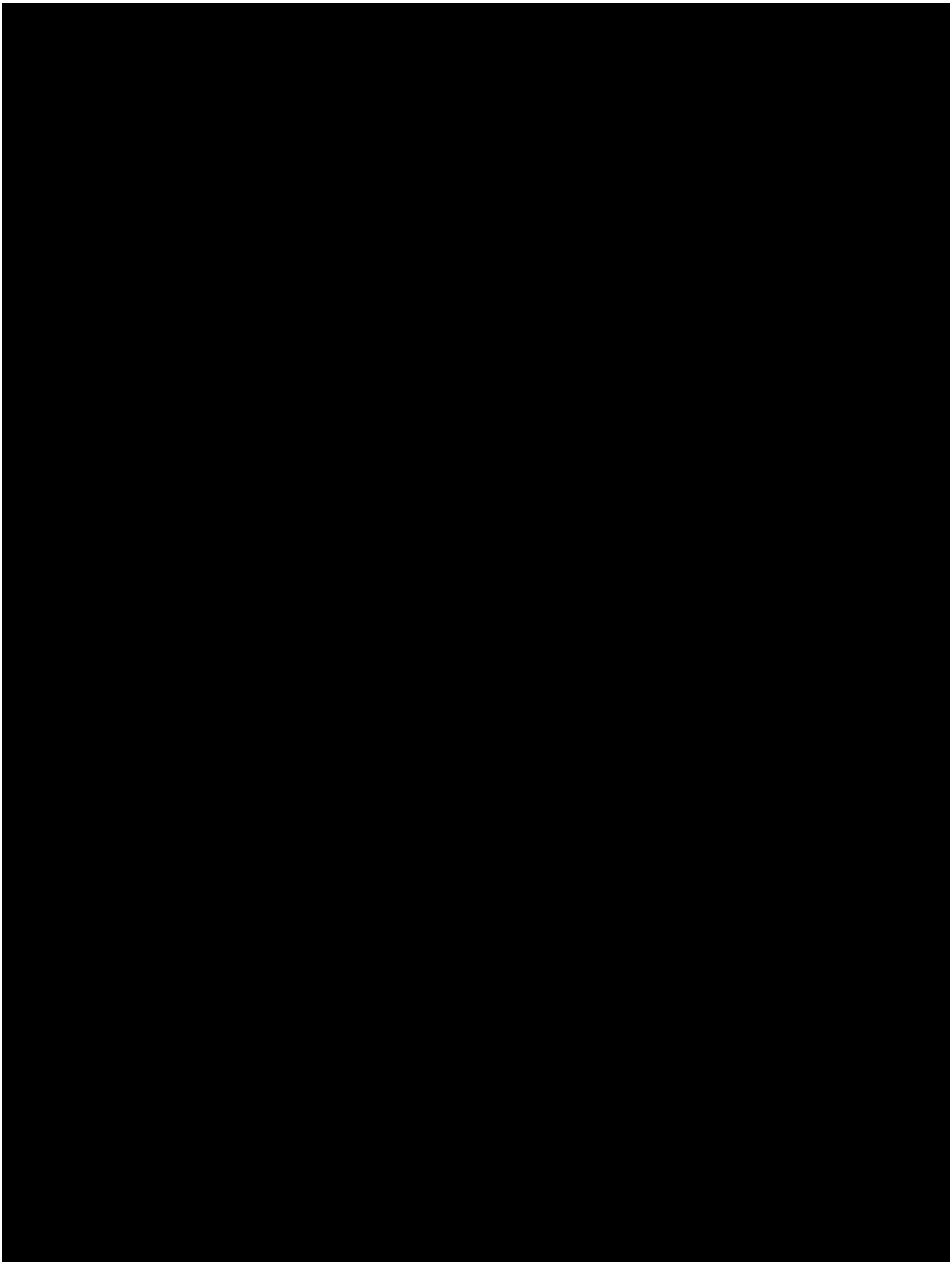


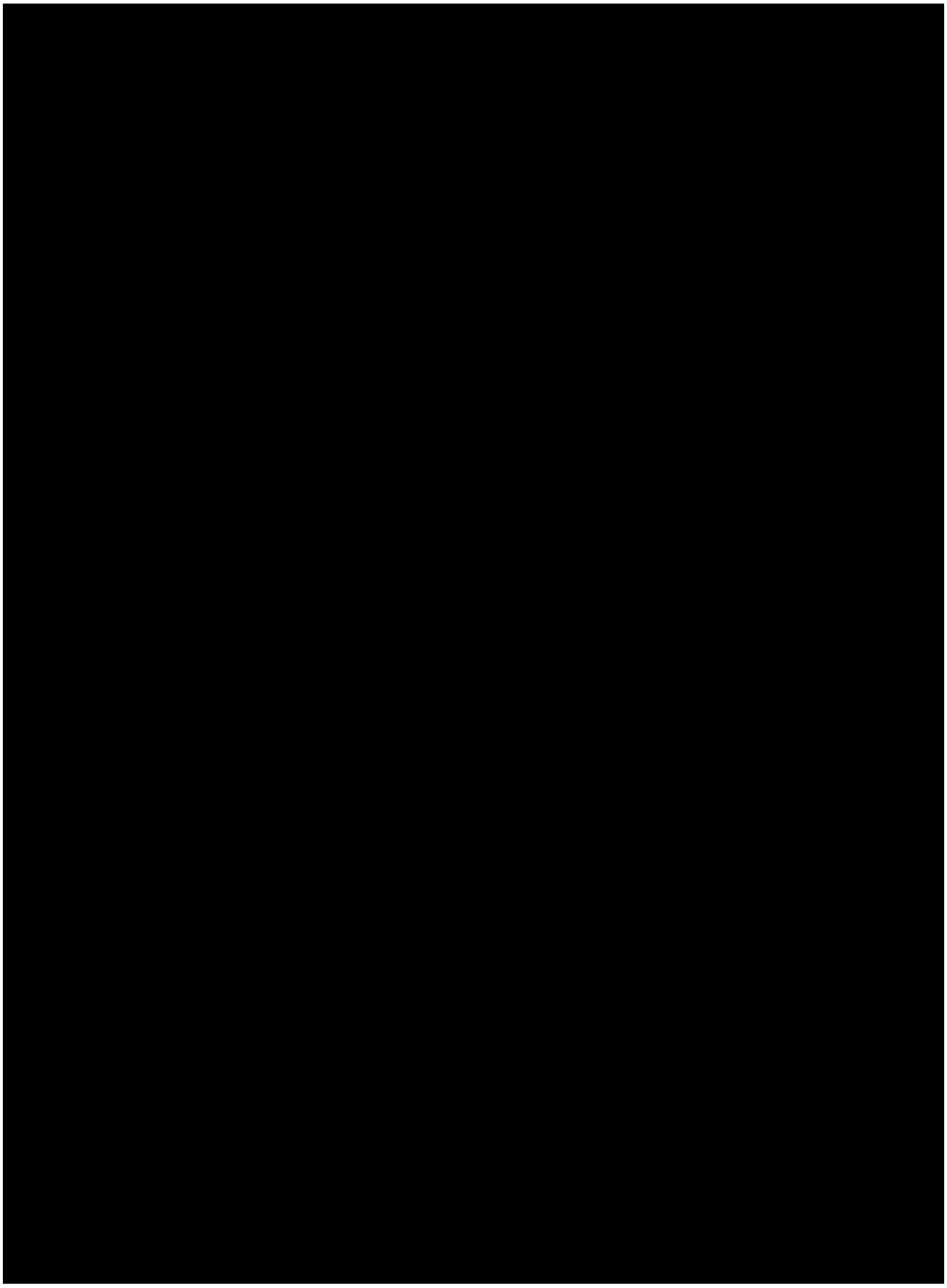


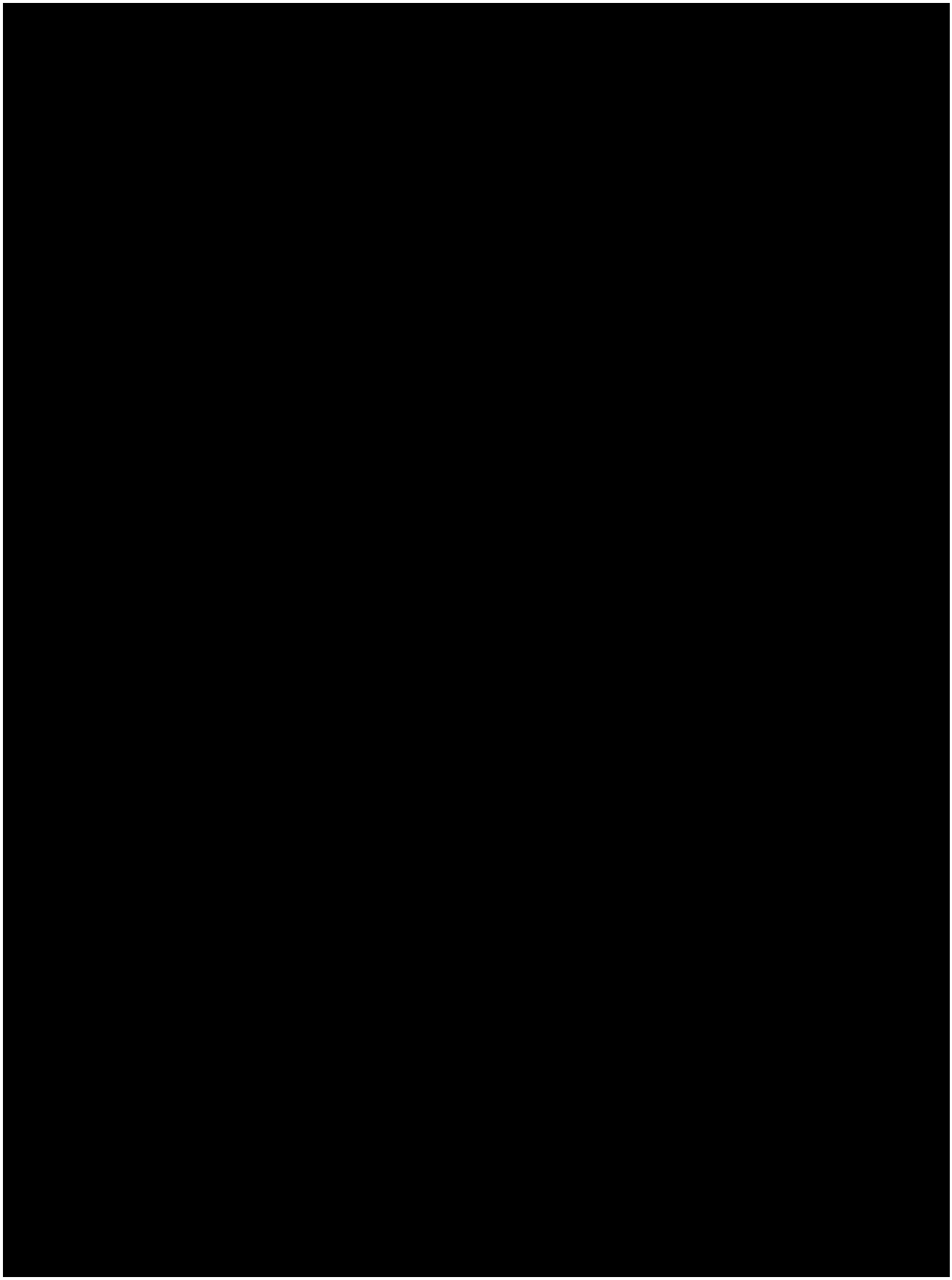


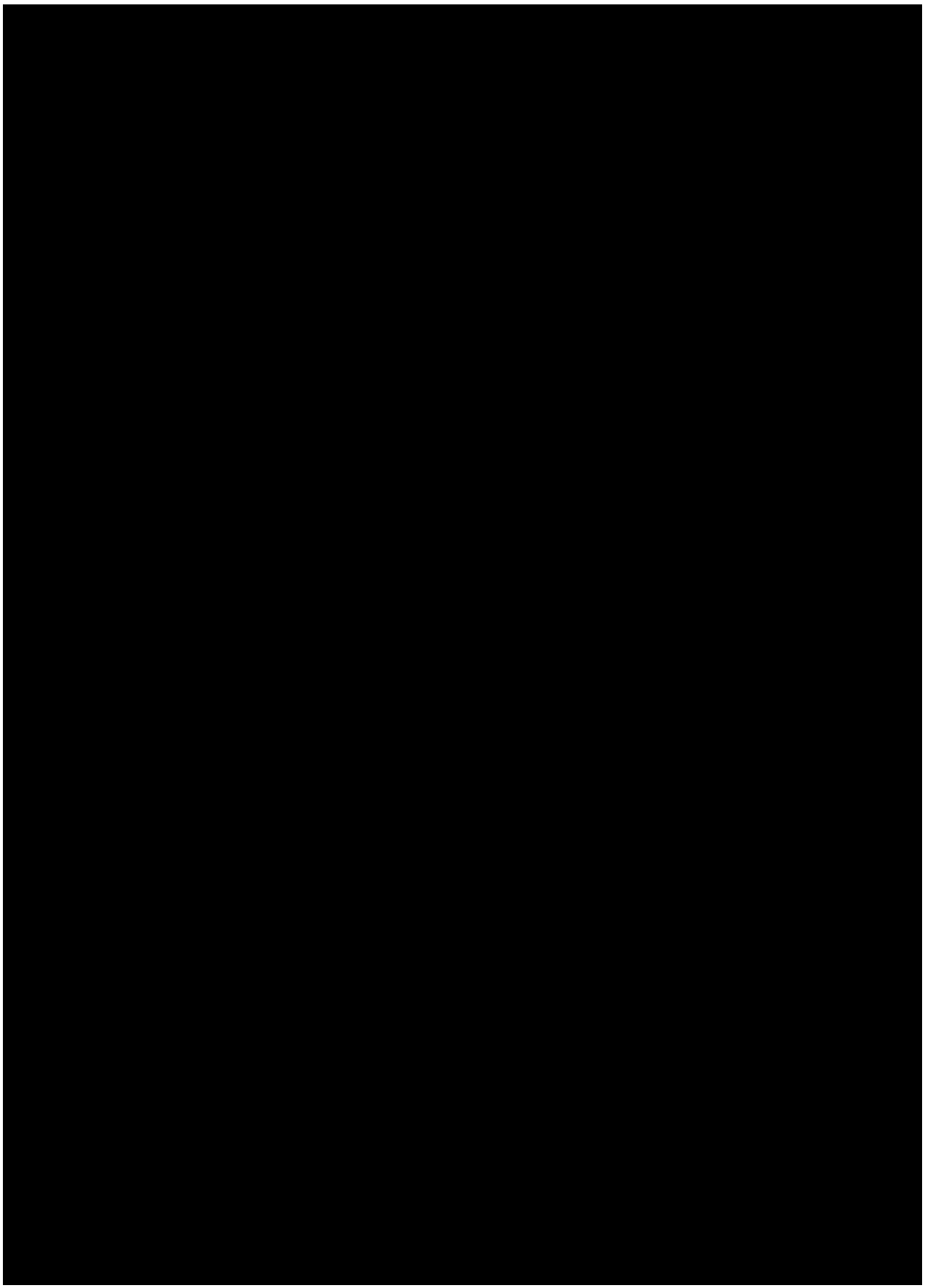












Chapter 4

Diagnostic accuracy of structural and functional tests for the detection of sight-threatening eye diseases in a community setting

4.1. Background

The burden of sight loss disproportionally affects the elderly (Evans and Rowlands, 2004) with one in five people aged 75 or over and one in two people aged 90 and over suffering from visual impairment (RNIB 2009). Sight loss from cataract, chronic open angle glaucoma (COAG), age-related macular degeneration (AMD), diabetic retinopathy and uncorrected refractive error affects approximately 2 million of the United Kingdom (UK) population (Pezzulo et al, 2018). Current case-finding strategies have been shown to miss half of the population affected by conditions such as COAG (Tielsch et al, 1991; Klein et al, 1992; Mitchell et al, 1996; Quigley and Vitale, 1997; Wensor et al, 1998). Poor access to routine NHS sight testing in 'at risk' populations, coupled with the absence of symptoms in the early stages of many of these diseases can lead to delayed or late presentation (Boodhna and Crabb, 2015; Lane et al, 2015).

Given that a significant proportion of sight loss can be prevented through early detection and timely therapeutic interventions, it is likely that investment in prevention and screening for sight-threatening eye disease would lead to a reduction in the burden of visual disability (WHO 2013). Currently the only sight threatening eye disease that satisfies the Wilson and Jungner's criteria (Wilson and Jungner, 1968) for a screening program in the UK is diabetic retinopathy, conditions such as glaucoma or macular degeneration, although fulfilling most criteria are not considered to be cost effective. (Hopley et al, 2004; Burr et al, 2007).

Earlier studies have investigated the diagnostic accuracy of individual screening tests to detect sight-threatening eye disease in general or clinic-based populations and established that no single test has sufficient accuracy (Ariyasu et al, 1996; Wang et al, 1998; Ivers et al, 2001; Boland et al, 2016). Recent studies have shown that by combining screening tests diagnostic performance can be significantly improved. For example, Kopplin and Mansberger demonstrated that a battery of tests of ocular structure and function performed by ophthalmic technicians were

effective in the detection of visually significant eye disease in a population of American Indian and Alaskan Native participants (Kopplin and Mansberger, 2015).

The primary objective of this study was to determine the diagnostic accuracy of modern imaging and visual function testing technologies, used alone and in combination, for detecting chronic open angle glaucoma (COAG) compared to a reference standard ophthalmic examination that included standard automated perimetry (SAP). We also conducted an exploratory secondary analysis to determine the performance of combinations of the screening tests for detecting other sight-threatening eye diseases.

4.2. Methods

Recruitment of participants

This prospective diagnostic accuracy study was conducted in a University community eye clinic in London between September 2012 and September 2013. Men and women aged ≥60yrs were invited to participate via a written invitation sent to community groups and local optometry practices. There were no exclusion criteria and participants with pre-diagnosed ocular disease were included. Participants underwent a series of technology-based index tests followed by a reference standard ophthalmic examination, conducted on the same day, to establish ocular health status. The study was approved by the School of Health Sciences Research and Ethics Committee, City, University of London and adhered to the tenets of the Declaration of Helsinki. Informed written consent was obtained prior to participation. All subjects underwent a series of technology-based index tests carried out by the author (BF), followed by a reference standard ophthalmic examination, conducted by an experienced clinician (PD), who had been independently validated in glaucoma and medical retina by sub-specialist ophthalmologists at Moorfields Eye Hospital.

The study was designed and reported in accordance with the Standards for Reporting of Diagnostic Accuracy (STARD) guideline. Figure 4.1. shows the study flow diagram.

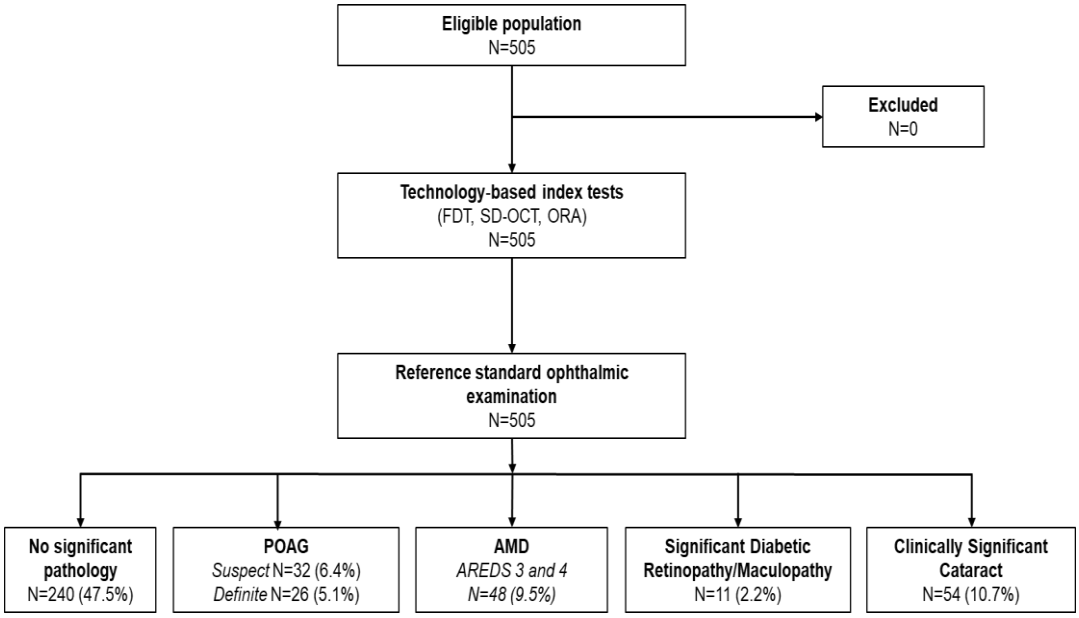


Figure 4.1. Study Flow Diagram. The number of eye conditions exceeds the number of participants due to co-morbidity. *Abbreviations: AREDS=Age-related Eye Disease Study; FDT=Frequency Doubling Technology Perimetry; SD-OCT=Spectral Domain Optical Coherence Tomography; ORA=Ocular Response Analyzer; AMD=Age-related macular degeneration: COAG=Chronic Open Angle Glaucoma.*

Description of index tests

Frequency Doubling Technology perimeter

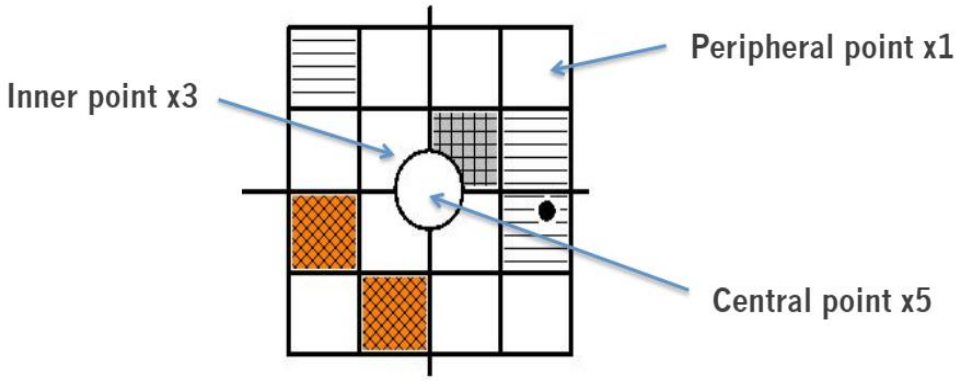
The frequency doubling illusion was initially reported by Kelly in 1966 (Kelly, 1966). The frequency doubling effect occurs when a low spatial frequency sinusoidal grating undergoes high temporal frequency flicker in counter phase. This causes the grating to appear to be twice its actual spatial frequency. Frequency doubled stimuli were incorporated in the first generation 'Frequency Doubling Technology (FDT) perimeter in 1997, as a means of screening for glaucomatous visual field defects (Johnson and Samuels, 1997). This first-generation instrument used 10° targets and incorporated both supra-threshold and threshold screening algorithms.

The C20-5 supra-threshold algorithm initially presents stimuli at a contrast level that can be detected by 95% of the normal age matched population. If the stimuli are seen, the test locations are classified as within normal limits. Targets missed at the 95% level are then retested at contrast levels that 98% and 99% of the age-marched normal population can detect. The depth of any defect can therefore be classified as 'mild' (significance level P<5%), 'moderate' (P<2%) or 'severe' (P<1%).

In the current study, a first generation FDT perimeter was used in the C20-5 supra-threshold mode. An abnormal result was defined using 2 cut-offs:

- ≥1 location(s) missed at the P < 5% significance level
- ≥1 location(s) missed at the P < 1% significance level

Further analysis of the FDT output can be performed using a scoring system described by Patel et al. (Patel et al., 2000). The algorithm allocates an overall score for each FDT result giving increased importance to more severe defects and locations missed closer to fixation (Patel et al, 2000). Figure 4.2 shows the point score for each stimulus location. The final score can be determined by adding scores for all missed points with scores ranging from 0 to 87.



Depth of defect	Significance level	Multiplying factor
Within normal limits	P ≥ 5%	0x
Mild relative loss	P < 5%	1x
Moderate relative loss	P < 2%	2x
Severe loss	P < 1%	3x

Figure 4.2. Patel et al., 2000 scoring algorithm of FDT supra-threshold results

Moorfields Motion Displacement Test

The Moorfields Motion Displacement Test (MMDT), which was developed collaboratively by Moorfields Eye Hospital, London, the UCL Institute of Ophthalmology and City, University of London, is a supra-threshold test for detecting sensitivity loss across the field of vision in glaucoma (Ong et al., 2014). The test uses 31 moving line stimuli that are displayed on a standard laptop computer at test locations corresponding to the 24-2 pattern of the Humphrey Visual Field Analyser (HFA). Peripheral stimuli are scaled in size by estimates of

retinal ganglion cell density, and with respect to age and eccentricity. The subject is positioned at 30cm using a purpose-built chin rest and stand.

Each line is subjected to a brief period of horizontal oscillation at a frequency of approximately 5 Hz in a random sequence. Whilst maintaining central fixation, the observer is asked to indicate whenever they detect any line movement. The subject's responses are then compared to a normative database and the results are recorded on a pass-fail probability plot (Figure 4.3) that provides an estimate of the 'probability of true damage' (PTD) at each test location between 0 and 100. A higher global PTD representing a greater probability of visual field damage. In this study, an abnormal plot was defined by the developers' recommended threshold of a global PTD≥3.0.

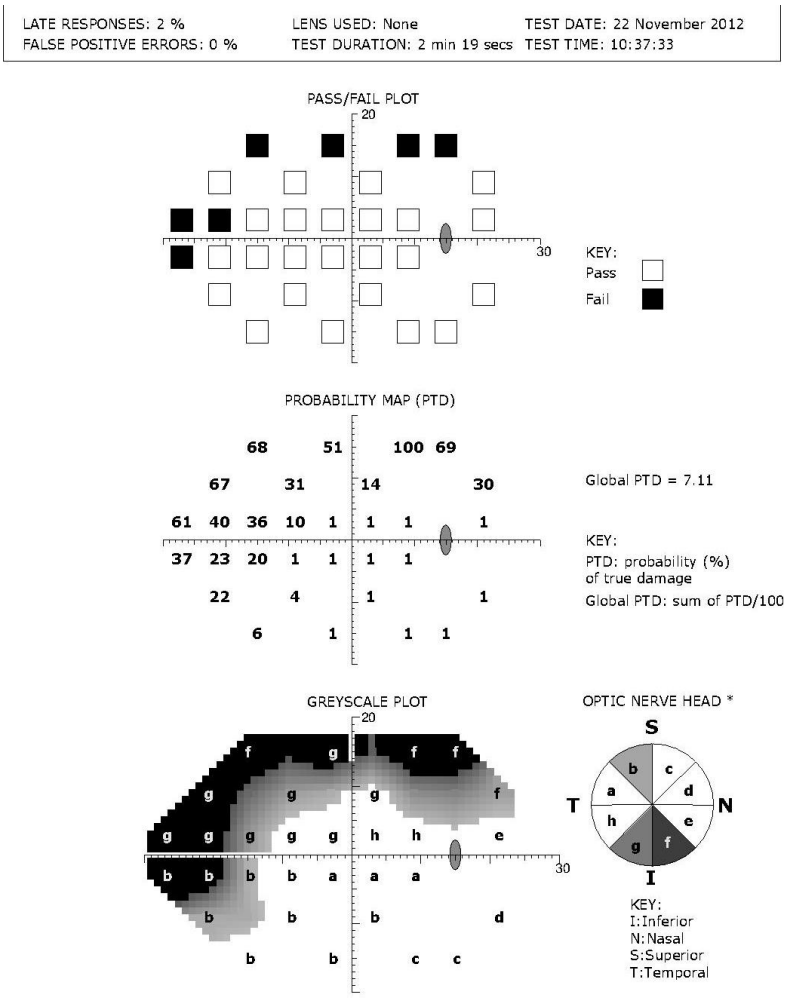


Figure 4.3. Moorfields Motion Displacement Threshold test (MMDT) output

IVue Spectral domain OCT

Optical coherence tomography (OCT) is an imaging technique that uses the principle of low-coherence interferometry to generate high resolution cross-sectional images of ocular structures. The development of Spectral or Fourier domain (SD) technology allowed faster image acquisition, higher image resolution, and improved retinal layer segmentation compared with the previous time-domain systems (TD) (Schuman, 2008).

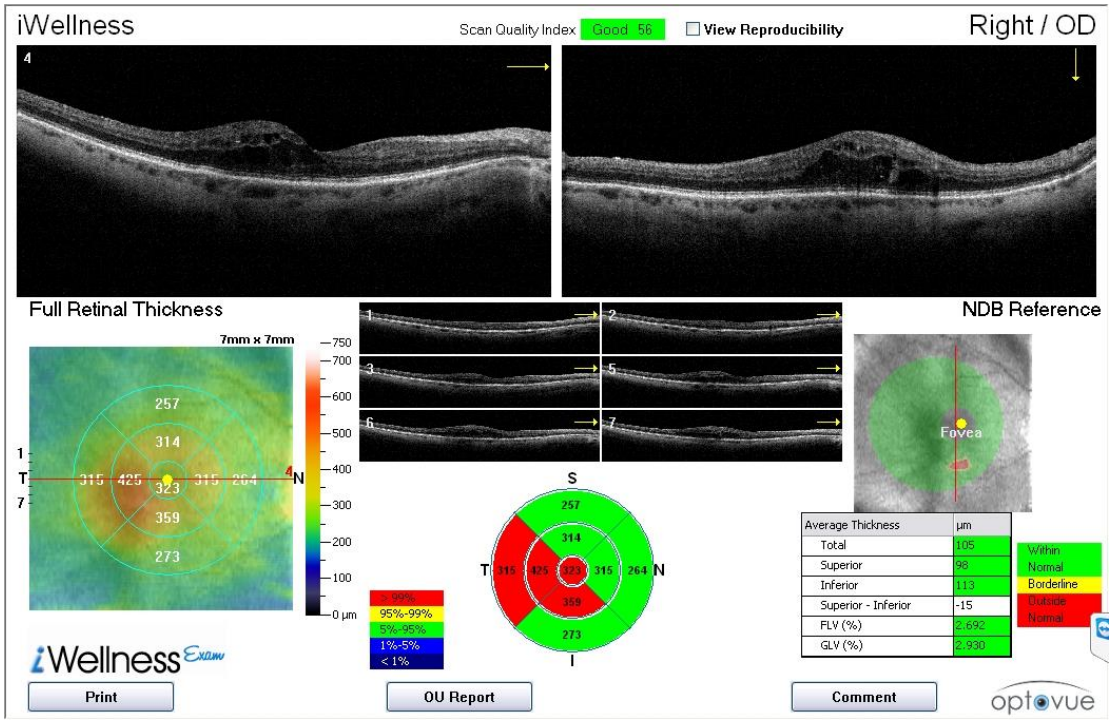
The iVue spectral domain OCT (SD-OCT) is a smaller, portable version of the larger RTVue OCT that provides a range of retinal and optic nerve scans with normative database comparisons (Figure 4.4). For the current study, we extracted data from the following scans:

- Nerve Fibre Optic nerve head (ONH) scan provides retinal nerve fibre layer (RNFL) thickness measurements from a circular area of 4.93mm radius from the disc centre
- iWellness screening protocol, which provides data on the integrity of the retina and optic nerve. The iWellness report provides 8 high resolution cross-sectional images along with quantitative data on full retinal thickness and ganglion cell complex (GCC) thickness.

A Scan Quality Index (SQI) below 40 generally represents a poor-quality scan, however the manufacturer advised that any decision to exclude data should also be based on a subjective evaluation of the scan. The first scan of the two captures was used for analysis unless it was excluded on the basis of poor quality (SQI <40 and/ or subjective evaluation). Scans were initially captured in low light conditions to encourage pupil dilation and without the use of mydriatic agents. If media opacities or miotic pupils precluded capture of adequate quality data, imaging was repeated following pupil dilation

The following structural parameters were included in the quantitative analysis (Table 4.1.)

Table 4.1. OCT structural parameters	
Parameter	**Description**
Optic nerve head peripapillary Retinal nerve fibre layer (ONH-RNFL)	The segmentation algorithm of the iVue OCT detects the internal limiting membrane and outer border of the RNFL in each A-scan. RNFL thickness is extracted from a peripapillary annulus of data points, 3.45mm from the centre of the optic disc, to construct a peripapillary RNFL map, which is divided into 8 segments. Each segment is color-coded based on a comparison of the average thickness compared to the age-matched normative database.
Ganglion cell complex (GCC)	The ganglion cell complex (GCC) comprises the ganglion cell layer, together with the adjacent RNFL and inner plexiform layers. GCC thickness data is acquired from a 7mm by 7mm square area that is centred on the fovea. These data are compared to the normative database reference values and reported as within normal limits, borderline or outside normal limits
Global loss volume (GLV)	Measures the average GCC loss over the entire GCC map.
Focal loss volume (FLV)	Measurements of the average amount of focal loss over the entire GCC map. The FLV detects focal loss using a pattern deviation map, similar to the corrected pattern standard deviation in visual fields plots.
Full retinal thickness	Full retinal thickness is measured from the 7 x7 mm macular scan and is presented as a retinal map showing the average retinal thickness in the nine areas as defined by the Early Treatment Diabetic Retinopathy Study (ETDRS).



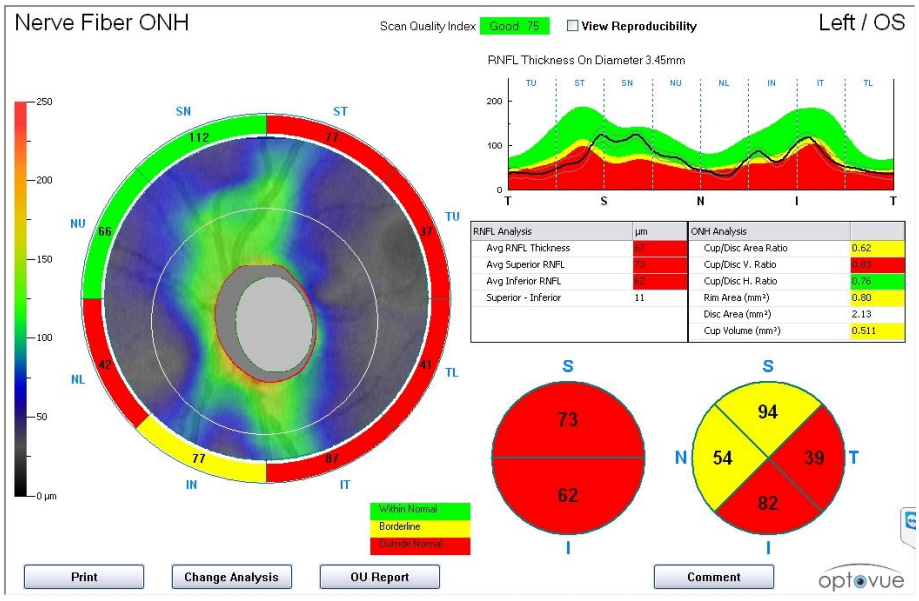


Figure 4.4. Vue SD-OCT scan outputs; iWellness (composite of GCC and retina map protocols) and ONH retinal nerve fibre layer thickness (RNFL)

Ocular response analyser

The Ocular Response Analyser (ORA) (Reichert Ophthalmic Instruments, Buffalo, NY) is a non-contact tonometer that measures the corneal response to indentation by a rapid air pulse. An alignment system positions an air tube over the central cornea and applies a 20ms collimated air pulse of increasing force to produce progressive corneal deformation. The cornea passes through an inward

applanation state (P1) to indentation, followed by an outward applanation state (P2), before returning to a normal corneal curvature (Figure 4.5). The air pulse force at P1 and P2 is used to calculate four parameters:

- Two measures of corneal biomechanics; corneal hysteresis (CH) and corneal resistance factor (CRF). The former quantifies the viscoelastic mechanical damping ability of the cornea and the latter is thought to represent the corneas overall viscoelastic resistance.
- Goldmann-correlated IOP (IOPg) and corneal-compensated IOP (IOPcc). IOPg is analogous to standard NCT IOP measurements, whereas IOPcc is an estimate that uses a mathematical correction to minimize corneal dependence of IOP and therefore provides a better indication of the true IOP (Lau and Pye, 2011).

A minimum of 4 measurements from each eye was taken.

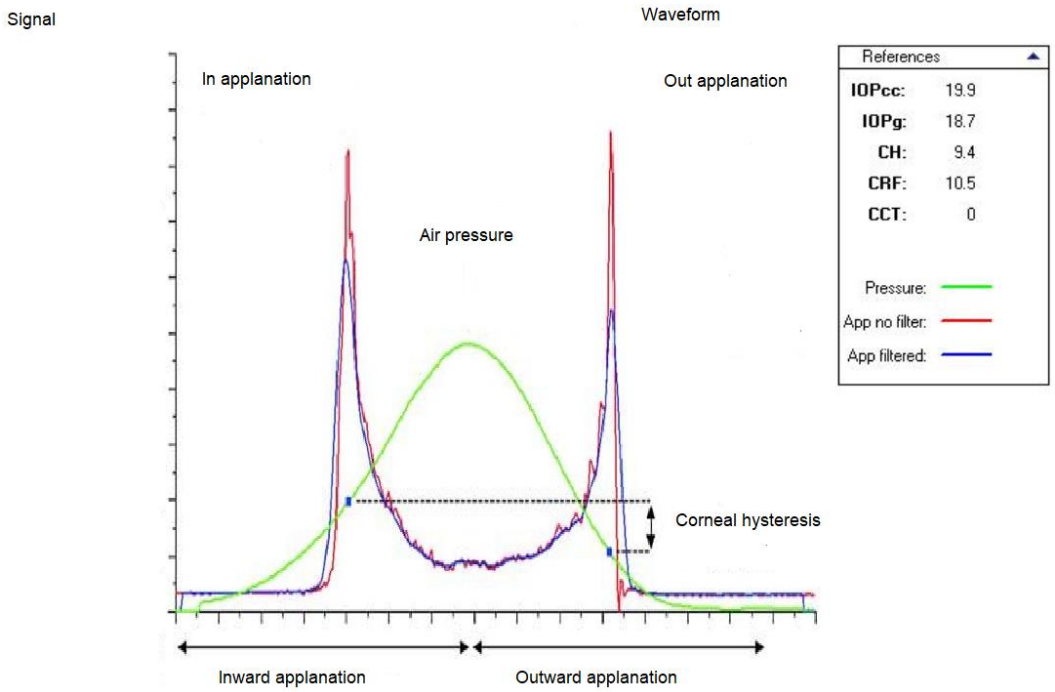


Figure 4.5. Ocular Response Analyser (ORA) graphical plot comprising 3 curves: pressure of air applied to the cornea (green), raw signal of applanation detection system (red), and a filtered version of the latter to identify optimum points of applanation (blue)

Index test quality indicators and defined thresholds

Quality/reliability indicators and thresholds to define abnormality were pre-defined for each index test (Table 4.2)

Table 4.2. A summary of index test quality indicators and cut-offs used in the study

Test	Indicators of suitable quality data	Cut-off / threshold for detection of COAG / suspected COAG
FDT perimeter	False positives (FP) < 15% Fixation errors (FE) < 15%	≥ 1 location missed at 5% level ≥ 1 location missed at 1% level
MMDT	False positives (FP) < 15% Late responses (LR) < 15%	Global PTD ≥ 3.0
iVue SD-OCT	Scan quality index (SQI) ≥ 40 and subjective evaluation of scans	P < 1% as defined by the normative database
ORA	Waveform score (WS) ≥ 6.5	Corneal hysteresis (CH) < 9.1 Corneal resistance factor (CRF) < 9.0

Reference standard examination

The reference standard examination was conducted on the same day as the index tests and the clinician was masked to the index test results.

The reference standard ophthalmic examination included:

- Full ophthalmic and medical history (with positive family history of glaucoma recorded if the subject reported a first-degree relative diagnosed with glaucoma)
- Measurement of visual acuity (using a 3m logMAR chart)
- Anterior segment assessment by biomicroscope (with eyes that had a potentially occludable angle identified by the van Herick test evaluated by gonioscopy and following pupil dilation the LOCS II system was used to grade nuclear, cortical, and subcapsular cataract)
- Measurement of IOP by the Goldmann applanation tonometer
- Dilated fundus examination (including a detailed optic disc examination) using indirect ophthalmoscopy with the slit-lamp
- Fundus photography (Topcon TRC-NW8F retinal camera)
- HFA used in 24-2 SITA standard mode (repeat testing was attempted for unreliable results and Glaucoma hemifield test (GHT) recordings of 'outside normal limits', either on the same day after a rest period or arranged for another day within a month of the study visit)

The following diagnostic criteria were used for classification of subjects as COAG and each of the potentially sight-threatening conditions that were included in the current analysis (Table 4.3).

Validation of the reference standard examiner

There is accumulative evidence that specialist optometrists, following additional training and accreditation, can make appropriate diagnostic and clinical management decisions when compared with a subspecialist ophthalmologist (Reeves et al 2016; Creer et al 2019; Harper et al 2020). The reference standard examiner for the current study was an experienced optometrist (PD) who undertook a clinical placement in glaucoma clinics at Moorfields Eye Hospital, London followed by an accreditation assessment. The accreditation assessment consisted of comparing diagnostic decisions with the classification of a consultant ophthalmologist (DGH or WN) on 50 patients. Each patient was classified as 'normal', 'suspect glaucoma', or 'glaucoma' based on the combined observation of the optic disc using binocular indirect ophthalmoscopy and visual field results. The

same classification system was used for further evaluation of the ability to classify glaucoma subjects by visual field assessment alone using 100-HFA C24-2 threshold field plots (50 right and 50 left eye plots). Results were compared with classification by a glaucoma consultant (DGH).

For accreditation in medical retinal conditions, the reference standard examiner attended the Moorfields Reading Centre for training and certification for grading ophthalmic images for diabetic retinopathy and age related macular degeneration. For training in diabetic retinopathy, the UK National Diabetic Retinopathy screening grading guidelines were used along with the Early Treatment of Diabetic Retinopathy Study standard images. PD initially graded the training images (minimum of 100 images of different quality and severity) and after achieving the required competency, was tested on batches of 25 sets of images to achieve the necessary accreditation. For Age Related Macular Degeneration, PD undertook training using the Wisconsin Age Related Macular Degeneration Study Folders, which show characteristic changes of all aspects of AMD stages. Once the examiner was in good agreement with the training set, she was tested using the accreditation set.

Table 4.3. Diagnostic criteria used for classifying the most prevalent eye diseases. Abbreviations: AREDS=Age-related Eye Disease Study; COAG=Chronic open angle glaucoma; LOCS II=Lens Opacities Classification System II.

Definitions of Eye Diseases		
Eye Condition	**Classification**	**Description**
COAG	Definite	1. Open anterior chamber angle 2. Localized absence of neuroretinal rim, cup-to-disc ratio ≥0.7, or interocular asymmetry in vertical cup-to-disc ratio ≥0.2 in similar sized discs 3. Presence of a concordant glaucomatous field defect based on criteria amended from Hodapp, Parrish and Anderson§
	Suspect	Features of glaucomatous optic neuropathy but with normal or equivocal fields or subjects with glaucomatous visual field defects but without concordant disc damage
Diabetic retinopathy*	Background retinopathy (R1)	Microaneurysms, dot blot haemorrhages, Venous loops, cotton wool spots
	Pre-proliferative retinopathy (R2)	Venous beading, venous reduplication, multiple blot haemorrhages, intraretinal microvascular abnormality (IRMA)
	Proliferative retinopathy (R3)	Active proliferative retinopathy
	Maculopathy (M1)	Groups of exudates, clinically significant macular oedema
Age-related Macular Degeneration (AMD)**	Early AMD (AREDS category 2)	Several small drusen or a few medium-sized drusen in one or both eyes
	Intermediate AMD (AREDS category 3)	Many medium-sized drusen or one or more large drusen (≥125μm), in one or both eyes
	Advanced AMD (AREDS category 4)	Geographic atrophy of the RPE involving the foveal center or any features of neovascular AMD
Clinically Significant Cataract	LOCS II grading***	LOCS II score ≥2.0 for cortex, posterior subcapsular, nuclear, or hypermature cataract

§ - Hodapp et al 1993 - Clinical decisions in glaucoma

* - The grading system for diabetic retinopathy was based on that used by the UK NHS Diabetic Eye Screening Programme. *https://www.gov.uk/government/publications/diabetic-eye-screening-retinal-image-grading-criteria*

** - The grading system for AMD was based on the classification system used in AREDS (Ferris et al 2005 - A Simplified Severity Scale for Age-Related Macular Degeneration: AREDS Report No. 18)

*** - Chylack et al 1989 - Lens opacities classification system II (LOCS II)

Index test acceptability questionnaire

Whilst the pupils were dilating, subjects were asked to complete a questionnaire regarding the acceptability of each of the index tests in terms of difficulty to perform, comfort and test duration using the form below (Figure 4.6). A full sized version of the questionnaire is available in Appendix 2.

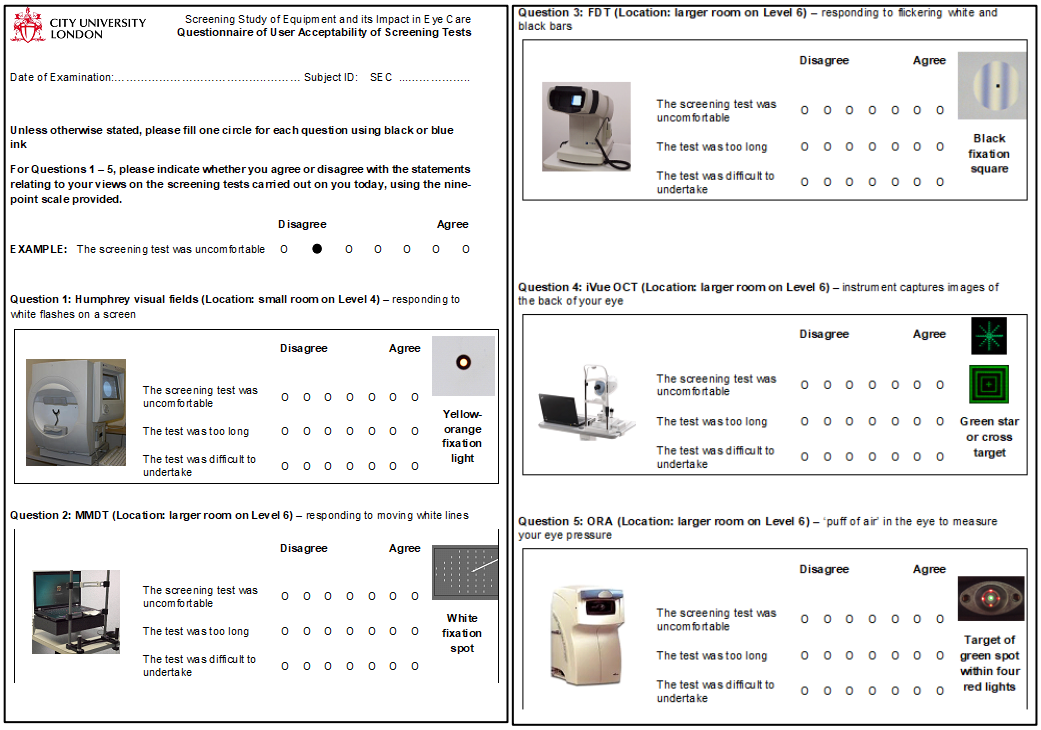


Figure 4.6. Index test acceptability questionnaire

Sample size calculation

The sample size was determined based on the primary outcome of the study, which was to determine the diagnostic accuracy of the index tests to detect COAG. This required an estimation of anticipated sensitivity of the tests, acceptable precision of the estimate and the prevalence of the condition in the local population. For a sensitivity of 0.75 (Mowatt et al, 2008), with a minimal acceptable precision of the sensitivity estimate of ±0.25 with 0.95 probability, we calculated that we would need approximately 50 cases of COAG. A previous cross-sectional study of an elderly population in North London (Reidy et al, 1998), found a prevalence of suspected and definite COAG of approximately 10%. We therefore estimated that 500 subjects needed to be recruited.

For the secondary analysis, we prospectively determined that a similar number of cases of intermediate and advanced AMD would be required, based on a disease prevalence of 8% in the target population (Reidy et al, 1998), anticipated index test sensitivity of 0.75 (Kopplin and Mansberger, 2015) with an equivalent level of precision.

Statistical analysis

Statistical analyses were performed using the following software: SPSS v21.0 (www.ibm.com/SPSS_Statistics), Medcalc v18.0 (www.medcalc.org) and STATA 13.0 (StataCorp LP, College Station, TX, www.stata.com). Unreliable results from visual function tests (FDT and MMDT) and repeatedly poor-quality data from the ORA and OCT were removed from the analysis.

Primary analysis: index test performance for the detection of COAG

For the initial analysis of test performance, summary statistics were tabulated and standard measurements of diagnostic accuracy (sensitivity, specificity, positive and negative predictive values) were calculated using 2 x 2 tables.

To compare index test performance within a clinically relevant range for detection of a low-prevalence disease, we determined the sensitivity at 90% specificity and normalised the partial area under the receiver operating characteristic curves (AUROC) to determine the average sensitivity between 90% and 100% specificity. To test for any statistically significant differences between sensitivities at a set specificity and differences in partial AUROC curve estimates, the Wald test was used (Pepe et al, 2009).

Secondary analysis: index test performance for the detection of sight-threatening eye disease

For the secondary analysis we adopted the statistical methods previously reported by Kopplin and Mansberger (Kopplin and Mansberger, 2015). Binomial logistic regression was used to identify univariate and multivariate associations between index test results and sight-threatening eye diseases. We also evaluated the predictive value of best-corrected visual acuity (<6/12). Those tests with univariate associations were included as potential covariates in a multivariate model using stepwise logistic regression. A significance level of <0.05 was set for entering and retaining each covariate in the final multivariate model. The diagnostic performance of the optimised screening panel was then calculated.

Analysis of user acceptability

Numerical data were grouped into summary tables based on the subject's response to the user acceptability survey, which asked them to rate agreement or disagreement with statements relating to particular characteristics of each test using a 7-point Likert scale. 'Disagree' was denoted by scores=1 to 3, 'neither agree nor disagree' was assigned to a score=4 and 'agree' for scores=5-7.

4.3. Summary of main results

Demographic and clinical characteristics of the study population

Table 4.4 describes the demographic characteristics of the 505 participants. This was a predominantly white population with proportionally more females than males (approx. 60:40). A significant proportion of the population suffered from diabetes (12.3%). A family history of glaucoma was reported by 16.4% of subjects.

Table 4.4. Demographic data for the 505 participants

	All subjects
N (%)	505 (100)
Age (years) **Median (IQR)**	68.0 ± 9 (59 to 77)
Gender (%) Male Female	 206 (40.8) 299 (59.2)
Ethnic group (%) White South Asian Black Chinese Other	 443 (87.7) 39 (7.7) 10 (2.0) 7 (1.4) 6 (1.2)
Positive family history of glaucoma (%)	83 (16.4)
Diabetes (type 1 or type 2) (%)	62 (12.3)

Table 4.5. provides a detailed summary of the ocular pathologies identified by the reference ophthalmic examination. Unsurprisingly, in this elderly population, a high proportion of participants had ocular morbidities. We adopted a strict definition for definitive COAG, which included both disc damage and a corresponding visual field defect. Twenty-six subjects (5.1%) fulfilled these criteria with a further 32 'COAG suspects' (6.4%), consisting of those with a suspicious disc or showing a field defect consistent with glaucoma but with an equivocal disc appearance. Of the 26 definite COAG cases, 11 (42%) were classified as early, 6 (23%) were classified as moderate, and 9 (35%) were classified as advanced using the criteria defined by Hoddap et al (Hoddap et al, 1993).

Other ocular pathologies identified in the cohort, which may or not have been mutually exclusive, included 9.5% subjects with intermediate or advanced AMD, 10.7% with clinically significant cataract in one or both eyes and 7.3% with diabetic retinopathy.

Table 4.5. Ocular pathology identified by the reference standard examination. Abbreviations: AMD=Age-related Macular Degeneration; COAG=Chronic open angle glaucoma; CHRPE= Congenital hypertrophy of retinal pigment epithelium.

Condition	**N**	**(%)**
AMD		
Early	155	(30.7)
Intermediate	27	(5.3)
Advanced	21	(4.2)
Diabetic Retinopathy		
Non-proliferative	26	(5.1)
Pre-proliferative	4	(0.8)
Proliferative	2	(0.4)
Clinically significant macular oedema	5	(1.0)
COAG		
Suspect	32	(6.4)
Definite	26	(5.1)
Ocular hypertension (OHT)	17	(3.4)
Cataract (clinically significant)	54	(10.7)
Other Retinal pathology		
Retinal detachment or tear (previous)	20	4.0
Choroiditis	3	0.6
Pigmented fundus lesion (naevus, CHRPE)	55	10.9
Chorioretinal atrophy/degeneration	25	5.0
Other retinal disorder	9	1.8
Other macular pathology		
Macular hole (lamellar or full thickness)	8	1.6
Epiretinal membrane: clinically significant	27	5.3
Other macular disorder	5	1.0
Other optic disc disorders	4	0.8
Corneal pathology	19	3.8
Corneal refractive surgery	9	1.8
Vitreal body opacity	10	2
Anterior segment disorder		
Primary angle closure suspect/ Primary angle closure	19	3.8
Pigment dispersion/ pseudo exfoliation	9	1.8
Uveitis (previous history)	2	0.4
Neurological disorder	6	1.2
Binocular vision disorder	37	7.3

Diagnostic accuracy of screening tests

Detection of COAG

Test performance was evaluated using the individual as the unit of analysis. Test thresholds at pre-defined cut-offs were compared for the most abnormal index test result from the right or left eye to the overall reference standard classification (Table 4.6).

Using an FDT threshold of one or more points missed at a P< 5% level of significance yielded a sensitivity of 92.3% and specificity of 65.2% for the detection of definite COAG. Test specificity improved to 79.1% using a criterion of one or more location(s) missed at a P < 1% with a similar level of sensitivity (88.5%).

The MMDT was evaluated using 2 different threshold criteria, using a PDT ≥ 2 as a cutoff threshold yielded a sensitivity of 69.2% and specificity of 77.5%, a second assessment with a PDT ≥ 3 resulted on a sensitivity of 65.4% and specificity of 81.2%.

We evaluated the performance of individual OCT parameters that fell outside the 99% normal limit for the detection of COAG (Table 4.6). All RNFL and GCC parameters performed well. The best performance was achieved using a criterion of an abnormality in any RNFL or GCC parameter, which yielded a sensitivity of 96.1% and specificity 81.3%.

A proportion of those with definite COAG had been previously diagnosed and were therefore already on ocular hypotensive therapy. Consequently, the IOP values recorded with the ORA had little diagnostic value for distinguishing those subjects with COAG from the rest of the sample.

The performance of index tests for detecting definitive and suspect COAG combined was variable (data not shown), with sensitivities ranging from 10.3% for the OCT (nasal quadrant) to 72.4% for the FDT (1 point missed at P<5% level).

Table 4.6. Sensitivity and specificity of index tests for detection of definite COAG		
Index test parameter	**Sensitivity (%) (CI)**	**Specificity (%) (CI)**
FDT 1 point missed at P<5% level	92.3 (75.9 - 97.9)	65.2 (60.8 – 69.3)
FDT 1 point missed at P<1% level	88.5 (71.0 – 96.0)	79.1 (75.2 – 82.5)
MMDT PDT ≥ 2	69.2 (50.0 – 83.5)	77.5 (73.5 – 81.0)
MMDT PDT ≥ 3	65.4 (46.2 – 80.6)	81.2 (77.5 – 84.5)
OCT Any RNFL	88.5 (71.0 – 96.0)	88.7 (85.5 – 91.2)
OCT Any GCC	80.8 (62.1 – 91.5)	87.9 (84.7 – 90.6)
OCT Any (GCC or RNFL)	96.1 (81.1 – 99.3)	81.3 (77.5 – 84.6)
ORA IOPg	19.2 (8.5 – 37.9)	88.9 (85.8 – 91.4)

Sensitivity at 90% specificity and partial AUROC curves for the range 90% to 100% specificity are summarized in Table 4.7

Table 4.7. Sensitivity at 90% specificity and Partial area under the receiver operation characteristic curve (AUROC) of index tests for detection of COAG		
Index test parameter	**Sensitivity at 90% specificity (%) (CI)**	**Partial AUROC from 90% to 100% Specificity (CI)**
FDT Patel score	61.5 (39.4 – 83.6)	0.35 (0.18 – 0.52)
MMDT Global PDT	55.7 (37.4 – 78.0)	0.44 (0.26 – 0.61)
OCT Mean RNFL	65.4 (47.1 – 83.7)	0.58 (0.41 – 0.76)
OCT RNFL Inferior quadrant	82.8 (67.6 – 97.9)	0.70 (0.53 – 0.86)
OCT Mean GCC	65.4 (47.1 – 83.7)	0.51 (0.34 – 0.67)
OCT GCC Inferior hemifield	69.2 (51.4 – 87.0)	0.61 (0.44 – 0.77)
OCT GCC – FLV	61.5 (42.4 – 80.7)	0.43 (0.27 – 0.59)
ORA IOPg	19.2 (3.9 – 34.6)	0.15 (0.02 – 0.27)

The inferior quadrant RNFL thickness was the best-performing OCT parameter, providing the highest sensitivity (82.8%) for the detection of COAG at 90% specificity and a partial AUROC of 0.70 from 90% to 100% specificity. The inferior quadrant RNFL thickness was statistically significantly superior to the FDT (Patel score) and MMDT (global PDT) based on partial AUROC curve analysis (Figure 4.7).

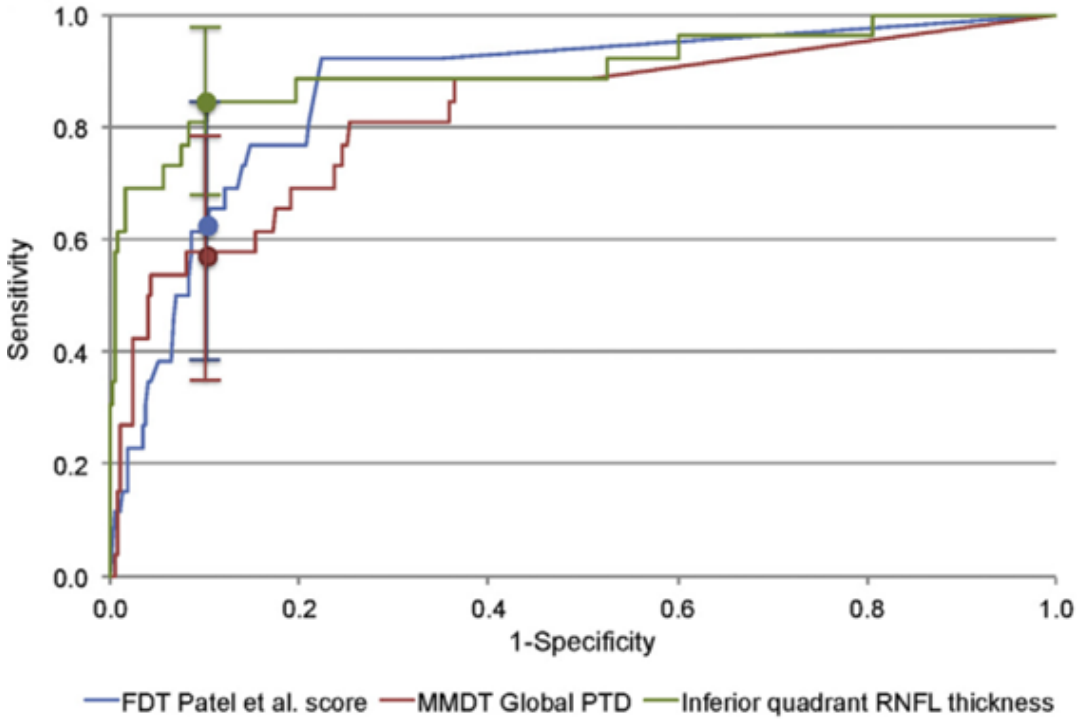


Figure 4.7. Index test diagnostic effectiveness comparisons using receiver operating characteristic curves with sensitivity at set specificity estimates and associated 95% confidence intervals for detection of COAG. FDT = Frequency Doubling Technology Perimeter; MMDT = Moorfields Motion Displacement Test; RNFL = retinal nerve fibre layer thickness.

Detection of any sight-threatening eye disease

Given the high prevalence of potentially sight-threatening eye disease in our elderly population, we conducted an exploratory secondary analysis to determine the predictive value of the index tests to identify any sight-threatening eye disease. For the purposes of this analysis sight-threatening eye disease was defined as: clinically significant cataract, suspect or definite COAG, intermediate or advanced AMD and significant diabetic retinopathy (see Table 4.3 for diagnostic definitions). In total 168 (33.5%) of the cohort met this definition. The performance of the individual screening tests for each of these conditions is shown in Table 4.8. Diagnostic precision of the individual tests was generally poor with sensitivities to detect any sight threatening disease ranging from 16.1% (IOP) to 61.9% (FDT <5%).

Univariate and multivariate logistic regression was used to determine the association between abnormal screening test results and the presence of sight-threatening eye disease. The impact of adding best-corrected visual acuity <6/12 from the reference test to the model was also evaluated and the diagnostic

performance of the final multivariate subset of screening tests was calculated. Table 4.9 shows the results of the multivariate regression analysis, which identified that visual acuity <6/12, abnormal FDT (≥1 point missed at 5% level) and peripapillary RNFL thickness outside the 99% normal limit were most predictive of any sight-threatening eye disease. The 3 screening tests had a sensitivity of 61.3% and specificity of 78.8% (Table 4.10).

Table 4.8. Sensitivity and Specificity of individual screening tests for each Eye Disease								
Screening test								
	Visual Acuity < 6/12		IOP		FDT 1% Level		FDT 5% Level	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Clinically significant cataract	42.6%	90.0%	16.7%	86.3%	46.3%	78.0%	64.8%	65.2%
AMD	37.5%	89.1%	12.5%	85.8%	45.8%	77.7%	60.4%	64.3%
COAG (Definite and suspect)	22.4%	87.7%	19.0%	86.6%	62.1%	80.3%	72.4%	66.4%
COAG (Definite)	38.5%	87.9%	19.2%	86.2%	88.5%	78.9%	92.3%	64.9%
Significant diabetic retinopathy	37.5%	86.9%	25.0%	86.1%	62.5%	76.1%	75.0%	62.6%
Any of the above	26.8%	93.2%	16.1%	86.9%	46.4%	86.4%	61.9%	73.9%
	SD-OCT (Full Retinal or GCC thickness)		SD-OCT (Full Retinal thickness)		SD-OCT (GCC thickness)		SD-OCT (Peripapillary RNFL thickness)	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Clinically significant cataract	42.6%	66.5%	35.2%	71.9%	25.9%	84.9%	13.0%	91.8%
AMD	52.1%	67.4%	50.0%	73.4%	37.5%	85.9%	25.0%	93.0%
COAG (Definite and suspect)	60.3%	68.9%	53.4%	74.4%	50.0%	88.1%	44.8%	96.0%
COAG (Definite)	76.9%	67.9%	65.4%	73.2%	80.8%	87.2%	69.2%	94.5%
Significant diabetic retinopathy	87.5%	66.4%	87.5%	72.1%	12.5%	83.6%	0.0%	91.1%
Any of the above	48.2%	72.5%	42.3%	77.9%	29.2%	90.1%	19.6%	96.7%

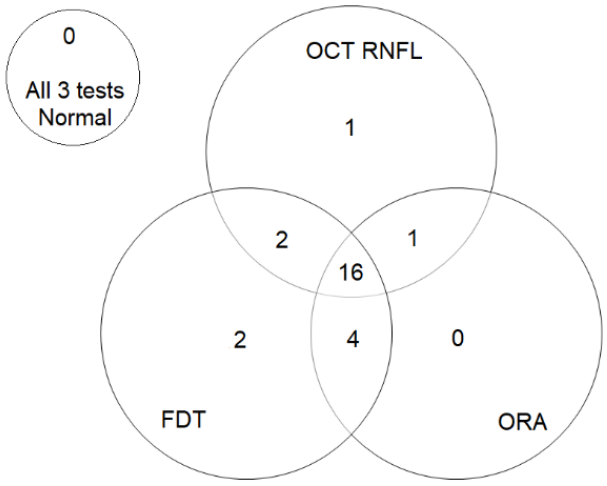
Table 4.9. Multivariate Stepwise Logistic Regression Analysis of Abnormal Screening Results with Ocular Disease. Abbreviations: CI -Confidence Interval, OR-Odds Ratio, FDT-Frequency-Doubling Technology Perimetry, COAG-Chronic Open Angle Glaucoma, IOP-Intra-Ocular Pressure, NS-Not Significant

	Clinically significant cataract	**COAG (Definite)**	**COAG (Definite and suspect)**	**Age related macular degeneration (AMD)**	**Significant Diabetic retinopathy/ maculopathy**	**Any sight-threatening eye disease**
	OR (95% CI)	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**
Visual acuity <6/12	5.53 (2.91, 10.54)	NS	NS	4.49 (2.40, 8.41)	NS	4.12 (2.26, 7.52)
IOP>21mmHg	NS	NS	NS	NS	NS	NS
FDT ≥1 point missed at 1% level	NS	8.55 (2.22, 32.96)	3.93 (2.05, 7.53)	NS	NS	NS
FDT ≥1 point missed at 5% level	2.60 (1.40, 4.84)	NS	NS	NS	NS	3.62 (2.38, 5.51)
SD-OCT (GCC thickness)	NS	5.52 (1.57, 19.41)	NS	NS	NS	NS
SD-OCT (Full retinal thickness)	NS	NS	NS	NS	17.86 (2.18, 146.47)	NS
SD-OCT (Peripapillary RNFL thickness)	NS	9.10 (2.85, 28.96)	11.98 (5.61, 25.60)	3.81 (1.74, 8.35)	NS	5.24 (2.45, 11.24)

Table 4.10. Diagnostic performance of the optimised panel of screening tests (SD-OCT, FDT and measurement of visual acuity in identifying sight-threatening eye disease)	
	Value (95% CI)
Sensitivity	61.3% (53.5–68.7)
Specificity	78.8% (74–83.1)
Positive Likelihood Ratio (PLR)	2.9 (2.3–3.7)
Negative Likelihood Ratio (NLR)	0.5 (0.40–0.60)
Disease Prevalence	33.7% (29.6–39.1)
Positive Predictive Value (PPV)	59.5% (53.7–65.2)
Negative Predictive Value (NPV)	80.0% (76.6–83.0)
Overall Accuracy	72.9% (68.8–76.8)

Figure 4.8 shows Venn diagrams combining index tests for the detection of COAG (Figure 4.8A) and any sight-threatening condition (Figure 4.8B). All 26 subjects with definite COAG were identified by the combination of peripapillary inferior RNFL thickness outside 99% confidence interval and an abnormal FDT (1 or more points missed lat the 5% level). In terms of any sight-threatening disease, the combination of an abnormal FDT, RNFL thickness outside 99% confidence limits and VA <6/12 identified 121 (72%) of those affected by the most severe disease. In the 47 subjects where 3 tests were recorded as normal; these 11 were diagnosed with cataract, 14 with non-proliferative diabetic retinopathy, 8 with suspect COAG and 14 with intermediate AMD.

A



B

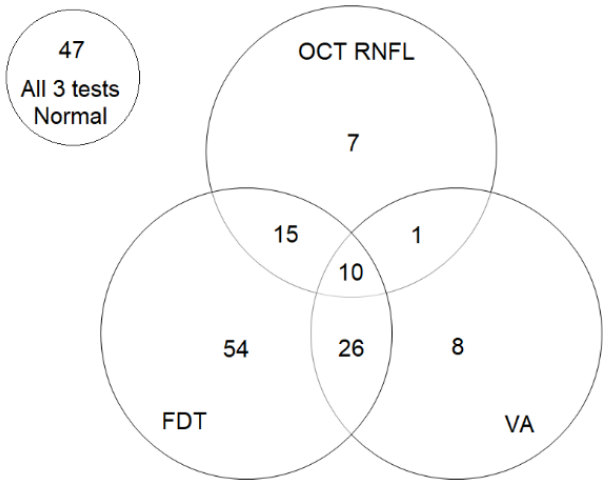


Figure 4.8. Venn diagrams presenting combined index test results for identification A) COAG and B) other sight threatening eye diseases.

Participant acceptability of screening tests

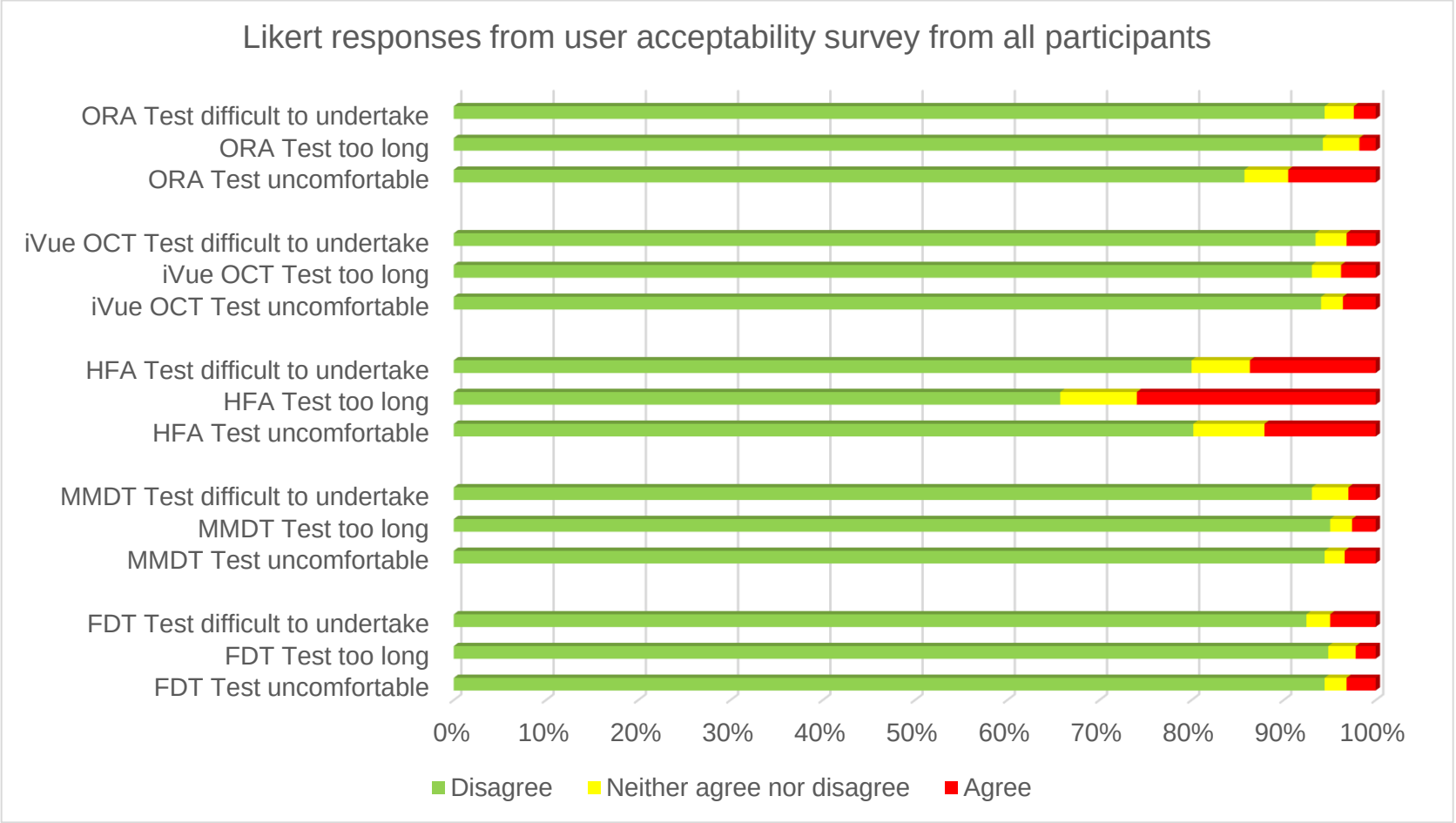


Figure 4.9. Aggregated Likert scale responses to index test acceptability survey in response to the statements a) 'Test was uncomfortable', b) 'Test was too long', and c) 'Test was difficult to undertake'

Figure 4.9 summarises aggregated Likert scale responses to the user acceptability survey. Overall, all index tests were well received with respondents finding that the tests were comfortable, not too long, and easy to perform. About 5% of respondents found that visual function tests (FDT and MMDT) were ‘uncomfortable’, ‘too long’ or ‘difficult to undertake’. By contrast, 14% and 26% of respondents found the HFA was difficult to undertake or too long respectively.

At the end of the survey 216 (43%) subjects responded to the ‘additional comments’ box. Responses were coded into three main categories:

- Responses relating to screening tests
- Responses relating to researchers
- Other comments.

Of the 139 comments relating to test experience, 74 (53%) were classified as ‘positive’ and 54 (39%) ‘negative’. The majority of the negative comments (65%, 24 of 37) made reference to the HFA followed by the FDT (19%, 7 of 37). General comments in the ‘negative’ group referred to tests being ‘tiring/ difficult’ (n=5), the need for concentration (n=10) and difficulty with posture during examinations (n=7)

4.4 Discussion

The overall aim of this study was to evaluate the diagnostic performance of modern structural and visual function screening tests for the detection of sight threatening eye conditions in a representative sample of elderly subjects, recruited from the community. The results of the screening tests, used alone or in combination, were evaluated against a reference standard ophthalmic examination, conducted by a trained clinician, who had been independently validated in the diagnosis of glaucoma and medical retinal conditions in the HES. We also investigated the acceptability to patients of each screening test to determine their suitability in this population.

Our initial objective was to determine the diagnostic accuracy of the technologies for glaucoma case-finding. Differentiating between COAG, suspect COAG and normal subjects presents a significant diagnostic challenge, due to the substantial overlap of clinical characteristics between these groups. There is also the problem of confounding due to other ocular pathologies that are likely to be present in an elderly population. Independent analyses were performed for those with manifest COAG (meeting a strict diagnostic definition based on disc appearance and corresponding glaucomatous field loss), and an analysis of a combined population of COAG cases and those with suspect COAG. Diagnostic accuracy was assessed using predefined cut-offs for abnormality and estimates of sensitivity, specificity, and likelihood ratios were calculated. We also derived estimates of sensitivity at 90% specificity and partial areas under the receiver operating characteristic curve (AUROC) from 90% to 100% specificity, to compare index test performance within a clinically relevant range for detection of a low-prevalence disease. Subjects were also asked to complete a questionnaire regarding the acceptability of each index test.

Overall, all index tests were well received with over 90% of respondents finding that the tests were comfortable, not too long, and easy to perform. The OCT was the most effective in identifying subjects with glaucoma. Using a criterion of any OCT parameter outside the 99% level, we would have identified 25 out of the 26 subjects diagnosed with definite COAG. In terms of individual OCT parameters, the inferior RNFL thickness showed the greatest diagnostic accuracy, with a sensitivity of 77% and specificity of 95%. Furthermore, inferior RNFL thickness was associated with a significantly greater partial AUROC than any of the visual-function tests. This probably reflects the increased vulnerability of the inferior

quadrant of the optic disc to glaucomatous damage (Jonas et al. 1993; Hood et al 2013).

All index tests showed poor discrimination between normal subjects and COAG suspects. Although we were able to show an improved sensitivity based on failure on either a structural or functional test, this was at the expense of a significantly reduced specificity. For case-finding of suspect COAG, we therefore propose a Bayesian strategy. This is based on the principal that in routine clinical practice, a clinician will intuitively integrate the results of several diagnostic tests together with a judgement of the patient's pre-test probability of COAG to determine likelihood of disease. The post-test probability can be formally calculated using widely available Bayesian diagnostic algorithms, which require the pre-test probability of disease and the likelihood ratios of the individual diagnostic tests used (Fagan 1975, Garway-Heath & Friedman 2003).

In 1968, Wilson and Jungner outlined ten criteria for appraising the viability, effectiveness and appropriateness of a screening programme (Wilson and Jungner, 1986). Although COAG fulfils criteria relating to the condition, availability of screening tests and treatment, in high-income countries, population screening for glaucoma is not considered to be cost-effective (Burr et al 2007; Moyer 2013). The cost-effectiveness of screening could be improved by combining screening tests to enhance diagnostic performance and also include more than one sight-threatening condition into the screening programme. We therefore conducted an exploratory reanalysis of our data to determine the performance of combinations of structural and functional screening tests to identify any sight-threatening eye disease. For the purposes of the current study we defined sight-threatening eye disease as: clinically significant cataract, actual or suspect COAG, intermediate or advanced AMD and significant diabetic retinopathy.

Using logistical regression, we established that the combination of reduced visual acuity (VA <6/12), abnormal FDT, and peripapillary RNFL thickness outside 99% normal limits had the best overall discriminatory power for the detection of any sight-threatening eye disease, with a sensitivity estimate of 61% and specificity 79%. The optimised test combination showed similarly high positive and negative predictive values to a previously published study conducted in Native Alaskans and American Indians (Kopplin and Mansberger 2015). The screening panel could form the basis of a screening model where the tests could be performed by a trained technician and screen positive individuals referred for further

investigation. Such a strategy could be particularly effectively in underserved populations with poor access to eye care.

In conclusion, this study provides useful preliminary data to inform the development of further larger, multicentre screening studies to validate this screening panel.

Strengths and limitations of the study

This study is the first of its kind in the UK to evaluate the performance of a combination of screening tests to detect clinically significant eye diseases in a primary care setting.

The major strengths in the study are:

- The design, analysis, and reporting complied with the STARD guidelines (Standards for Reporting of Diagnostic Accuracy)
- The target population included consecutive subjects who met the inclusion criteria and there were no exclusions, this was intended to reduce spectrum bias.
- The prevalence of sight-threatening eye diseases in our population was calculated around 30%, this value was similar to a London-based cross-sectional study that used random sampling (Reidy et al, 1998).
- The reference standard used to classify the participants ocular status corresponded to that used in a typical hospital eye clinic and was based on the results of a standard ophthalmic examination by a validated clinician.
- The reference standard examination and all index tests were performed on the same day and the clinician performing the reference standard examination and the ophthalmic technician undertaking the index tests were masked to each other results.

The study limitations:

- Although the population included consecutive subjects who met the inclusion criteria it is possible that higher numbers of those with a personal or family history of eye disease were more likely to agree to participate in the study.
- The sample size of 505 subjects provided between 2% to 10% of subjects with a specific sight threatening eye condition. The small number of

patients with those conditions resulted in wide CIs around our diagnostic sensitivity estimates.

- Roughly 90% of our study population was of white European origin and therefore our findings may not be generalizable to other ethnic groups.
- The current study did not include a formal cost-effectiveness analysis
- We used a pragmatic diagnostic reference standard, similar to the standard expected in specialist glaucoma and medical retina clinics in the UK. To this end, the reference examiner undertook an extensive process of training and accreditation using a standard methodology to ensure that they reached the required standard. A definitive reference standard would have required a dedicated Reading Centre with trained and accredited graders, similar to that used in clinical trials in ophthalmology.

4.5. Role in the study

My role in this study included conducting all index tests (FDT, MMDT, OCT and ORA) on the 505 participants. Another PhD student (Priya Dabasia) carried out the reference standard ophthalmic examination. I established the database for storage of index and reference test results and combined the data for the analysis. Working with Priya, I was involved in cleaning and checking the data and as part of the research team, contributed to data analysis and interpretation. For the publication of the results of the primary analysis 'Diagnostic Accuracy of Technologies for Glaucoma Case-Finding in a Community Setting (Ophthalmology 2015; 122:2407-2415) Priya drafted the manuscript and the results of this study were included as part of her PhD (A study of the role of advanced technologies in glaucoma case-finding).

Under Professor Lawrenson's supervision I was involved in the conception, design, data analysis and interpretation for the secondary objective (role of advanced technology in the detection of sight-threatening eye disease in a UK community setting). The findings were reported via poster at EVER 2019 and subsequently published in BMJ Open Ophthalmology. I wrote the first draft of this paper and subsequent revisions following input from members of the research team.

My percentage contribution to this study: 70%

4.6. References

Ariyasu, R.G., Lee, P.P., Linton, K.P., LaBree, L.D., Azen, S.P., Siu, A.L., 1996. Sensitivity, Specificity, and Predictive Values of Screening Tests for Eye Conditions in a Clinic-based Population. Ophthalmology 103, 1751–1760.

Boland, M.V., Gupta, P., Ko, F., Zhao, D., Guallar, E., Friedman, D.S., 2016. Evaluation of Frequency-Doubling Technology Perimetry as a Means of Screening for Glaucoma and Other Eye Diseases Using the National Health and Nutrition Examination Survey. JAMA Ophthalmology 134, 57.

Boodhna, T., Crabb, D.P., 2015. Disease severity in newly diagnosed glaucoma patients with visual field loss: trends from more than a decade of data. Ophthalmic and Physiological Optics 35, 225–230.

Burr, J.M., Mowatt, G., Hernández, R., Siddiqui, M. a. R., Cook, J., Lourenco, T., Ramsay, C., Vale, L., Fraser, C., Azuara-Blanco, A., Deeks, J., Cairns, J., Wormald, R., McPherson, S., Rabindranath, K., Grant, A., 2007. The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation. Health Technology Assessment 11, iii–iv, ix–x, 1–190.

Chylack, L.T., Leske, M.C., McCarthy, D., Khu, P., Kashiwagi, T., Sperduto, R., 1989. Lens opacities classification system II (LOCS II). Archives of Ophthalmology 107, 991–997.

Creer, R.C., Roberts, S.A., Aslam, T.M., Balaskas, K., Chhabra, R., Mahmood, S., Turner, G.S., Harper, R.A., 2019. Treatment decisions of UK hospital optometrists and ophthalmologists in patients with nAMD: a vignette study. Ophthalmic and Physiological Optics 39, 432–440.

Ervin, A.-M., Boland, M.V., Myrowitz, E.H., Prince, J., Hawkins, B., Vollenweider, D., Ward, D., Suarez-Cuervo, C., Robinson, K.A., 2012. Screening for Glaucoma: Comparative Effectiveness, AHRQ Comparative Effectiveness Reviews. Agency for Healthcare Research and Quality (US), Rockville (MD).

Evans, B.J.W., Rowlands, G., 2004. Correctable visual impairment in older people: a major unmet need. Ophthalmic and Physiological Optics 24, 161–180.

Fagan, T., 1975. Nomogram for Bayes's Theorem. New England Journal of Medicine 293, 257–257.

Garway-Heath, D., Friedman, D., 2006. How Should Results from Clinical Tests Be Integrated into the Diagnostic Process? Ophthalmology 113, 1479–1480.

Harper, R.A., Gunn, P.J.G., Spry, P.G.D., Fenerty, C.H., Lawrenson, J.G., 2020. Care pathways for glaucoma detection and monitoring in the UK. Eye 34, 89–102.

Hodapp, E., Parrish, R.K., Anderson, D.R., 1993. Clinical decisions in glaucoma. Mosby, St. Louis, Mo.

Hood, D.C., Raza, A.S., de Moraes, C.G.V., Liebmann, J.M., Ritch, R., 2013. Glaucomatous damage of the macula. Progress in Retinal and Eye Research 32, 1–21.

Hopley, C., 2004. Cost utility of screening and treatment for early age related macular degeneration with zinc and antioxidants. British Journal of Ophthalmology 88, 450–454.

Ivers, R.Q., Macaskill, P., Cumming, R.G., Mitchell, P., 2001. Sensitivity and specificity of tests to detect eye disease in an older population. Ophthalmology 108, 968–975.

Johnson, C.A., Samuels, S.J., 1997. Screening for glaucomatous visual field loss with frequency-doubling perimetry. Investigative Ophthalmology and Visual Science 38, 413–425.

Jonas, J.B., Fernández, M.C., Stürmer, J., 1993. Pattern of Glaucomatous Neuroretinal Rim Loss. Ophthalmology 100, 63–68.

Kelly, D.H., 1966. Frequency Doubling in Visual Responses*†. Journal of the Optical Society of America 56, 1628.

Klein, B.E.K., Klein, R., Sponsel, W.E., Franke, T., Cantor, L.B., Martone, J., Menage, M.J., 1992. Prevalence of Glaucoma. Ophthalmology 99, 1499–1504.

Kopplin, L.J., Mansberger, S.L., 2015. Predictive Value of Screening Tests for Visually Significant Eye Disease. American Journal of Ophthalmology 160, 538-546.e3.

Lane, M., Mathewson, P., Sharma, H., Palmer, H., Nightingale, P., Shah, P., Tsaloumas, M., Denniston, A., 2015. Social deprivation as a risk factor for late presentation of proliferative diabetic retinopathy. Clinical Ophthalmology 347.

Lau, W., Pye, D., 2011. A Clinical Description of Ocular Response Analyzer Measurements. Investigative Opthalmology & Visual Science 52, 2911.

Mitchell, P., Smith, W., Attebo, K., Healey, P.R., 1996. Prevalence of Open-angle Glaucoma in Australia. Ophthalmology 103, 1661–1669.

Mowatt, G., Burr, J.M., Cook, J.A., Siddiqui, M.A.R., Ramsay, C., Fraser, C., Azuara-Blanco, A., Deeks, J.J., OAG Screening Project, 2008. Screening tests for detecting open-angle glaucoma: systematic review and meta-analysis. Investigative Ophthalmology and Visual Science 49, 5373–5385.

Moyer, V.A., U.S. Preventive Services Task Force, 2013. Screening for glaucoma: U.S. Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine 159, 484–489.

Ong, E.-L., Zheng, Y., Aung, T., Tan, L., Cheng, C.-Y., Wong, T.-Y., How, A., 2014. Performance of the Moorfields Motion Displacement Test for Identifying Eyes with Glaucoma. Ophthalmology 121, 88–92.

Patel, S.C., Friedman, D.S., Varadkar, P., Robin, A.L., 2000. Algorithm for interpreting the results of frequency doubling perimetry. American Journal of Ophthalmology 129, 323–327.

Pepe, M., Longton, G., Janes, H., 2009. Estimation and Comparison of Receiver Operating Characteristic Curves. Stata Journal 9, 1.

Pezzullo, L., Streatfeild, J., Simkiss, P., Shickle, D., 2018. The economic impact of sight loss and blindness in the UK adult population. BMC Health Services Research 18.

Quigley, H.A., Vitale, S., 1997. Models of open-angle glaucoma prevalence and incidence in the United States. Investigative Ophthalmology and Visual Science 38, 83–91.

Reeves, B.C., Scott, L.J., Taylor, J., Harding, S.P., Peto, T., Muldrew, A., Hogg, R.E., Wordsworth, S., Mills, N., O'Reilly, D., Rogers, C.A., Chakravarthy, U., 2016. 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 6, e010685.

Reidy, A., Minassian, D.C., Vafidis, G., Joseph, J., Farrow, S., Wu, J., Desai, P., Connolly, A., 1998. Prevalence of serious eye disease and visual impairment in a north London population: population based, cross sectional study. open Journal 316, 1643–1646.

RNIB (2009). Future sight loss UK (1): The economic impact of partial sight and blindness in the UK adult population Full report. [online] Available at: https://www.rnib.org.uk/sites/default/files/FSUK_Report.pdf[Accessed 08/01/20].

Schuman, J.S., 2008. Spectral domain optical coherence tomography for glaucoma (an AOS thesis). Transactions of the American Ophthalmological Society 106, 426–458.

Tielsch, J.M., Katz, J., Singh, K., Quigley, H.A., Gottsch, J.D., Javitt, J., Sommer, A., 1991. A Population-based Evaluation of Glaucoma Screening: The Baltimore Eye Survey. American Journal of Epidemiology 134, 1102–1110.

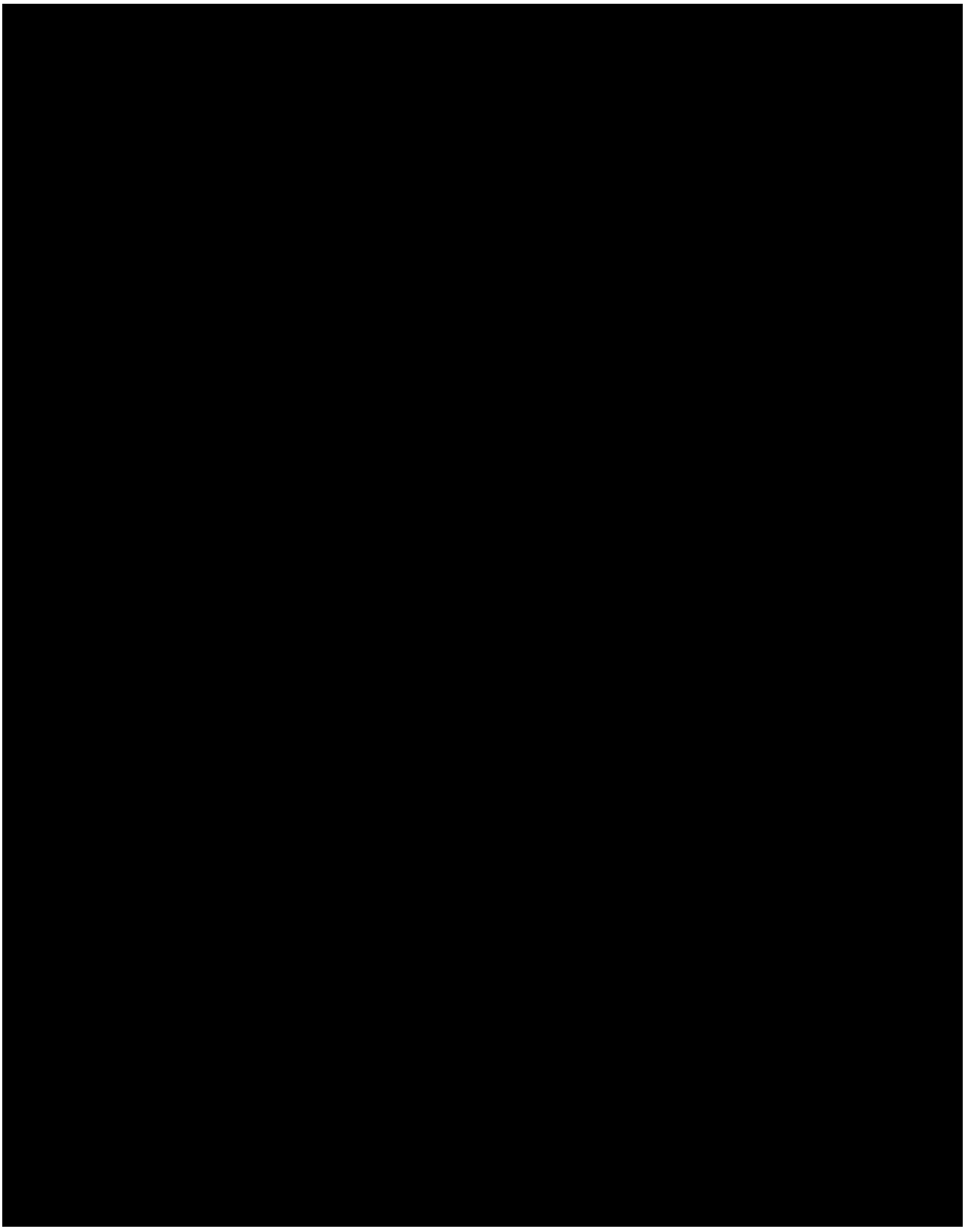
Wang, F., Tielsch, J.M., Ford, D.E., Quigley, H.A., Whelton, P.K., 1998. Evaluation of screening schemes for eye disease in a primary care setting. Ophthalmic Epidemiology 5, 69–82.

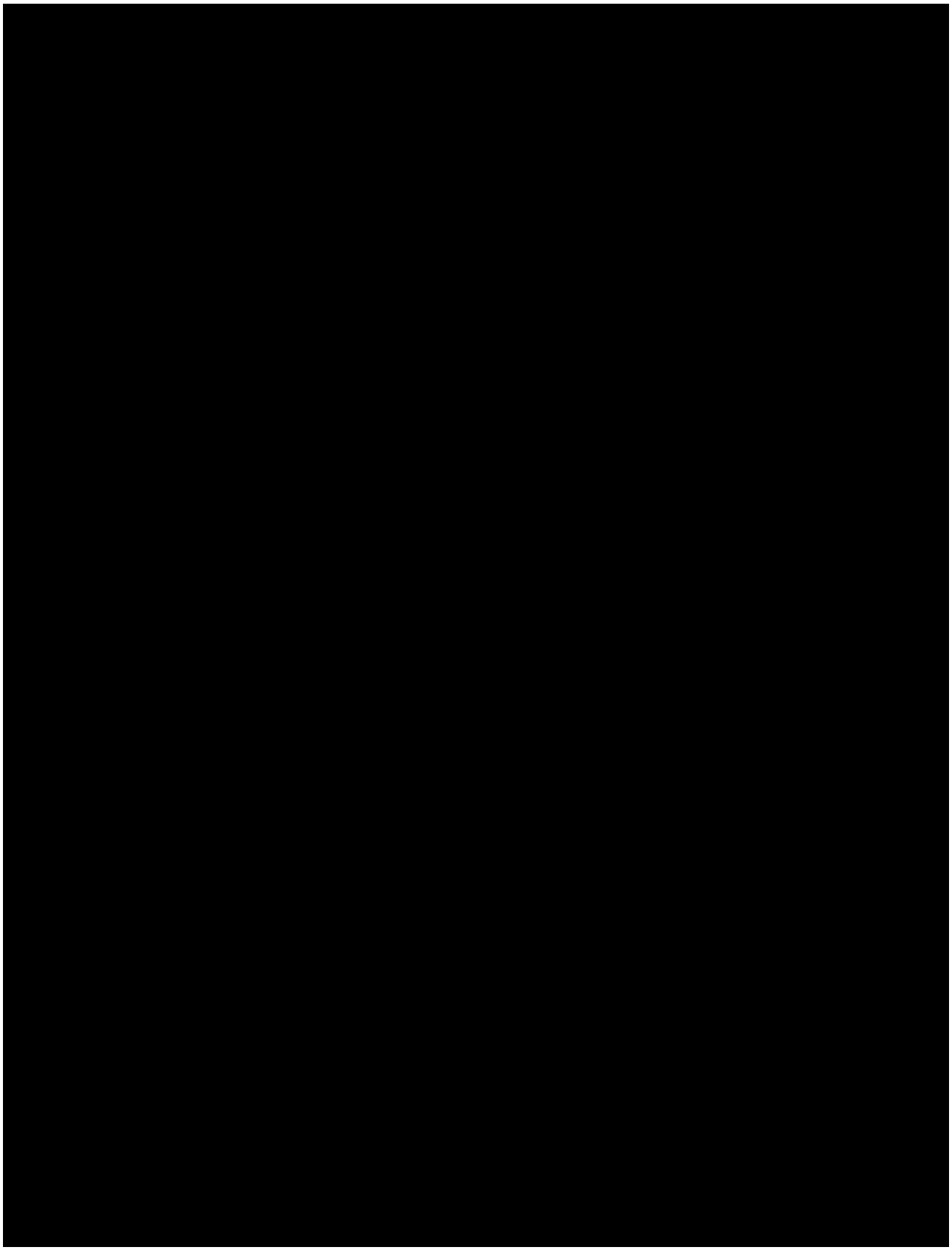
Wensor, M.D., McCarty, C.A., Stanislavsky, Y.L., Livingston, P.M., Taylor, H.R., 1998. The prevalence of glaucoma in the Melbourne visual impairment project. Ophthalmology 105, 733–739.

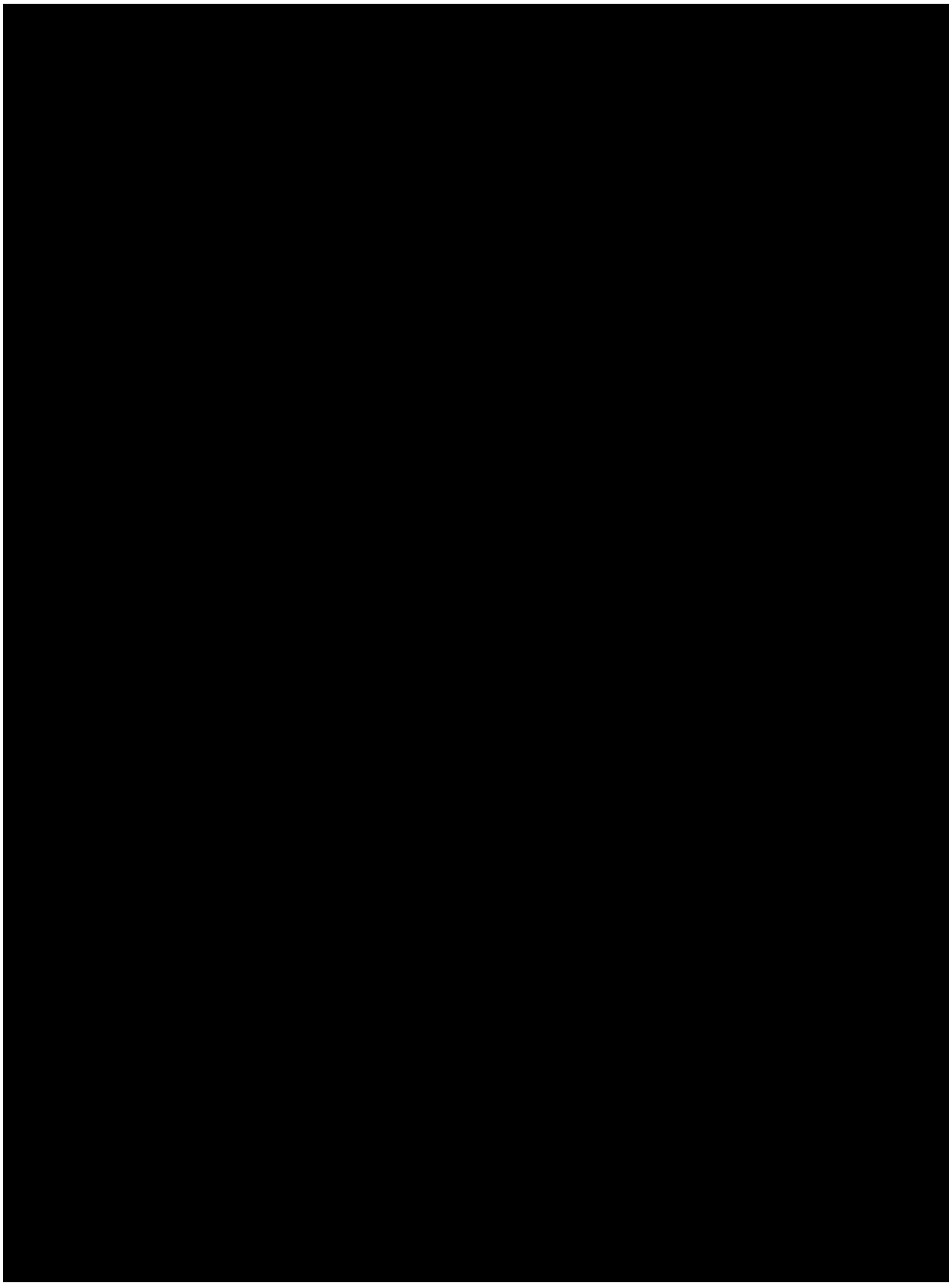
Wilson, J.M.G., Jungner, G., 1968. Principles and Practice of Screening for Diseases [by] J.M.G. Wilson [and] G. Jungner. World Health Organization.

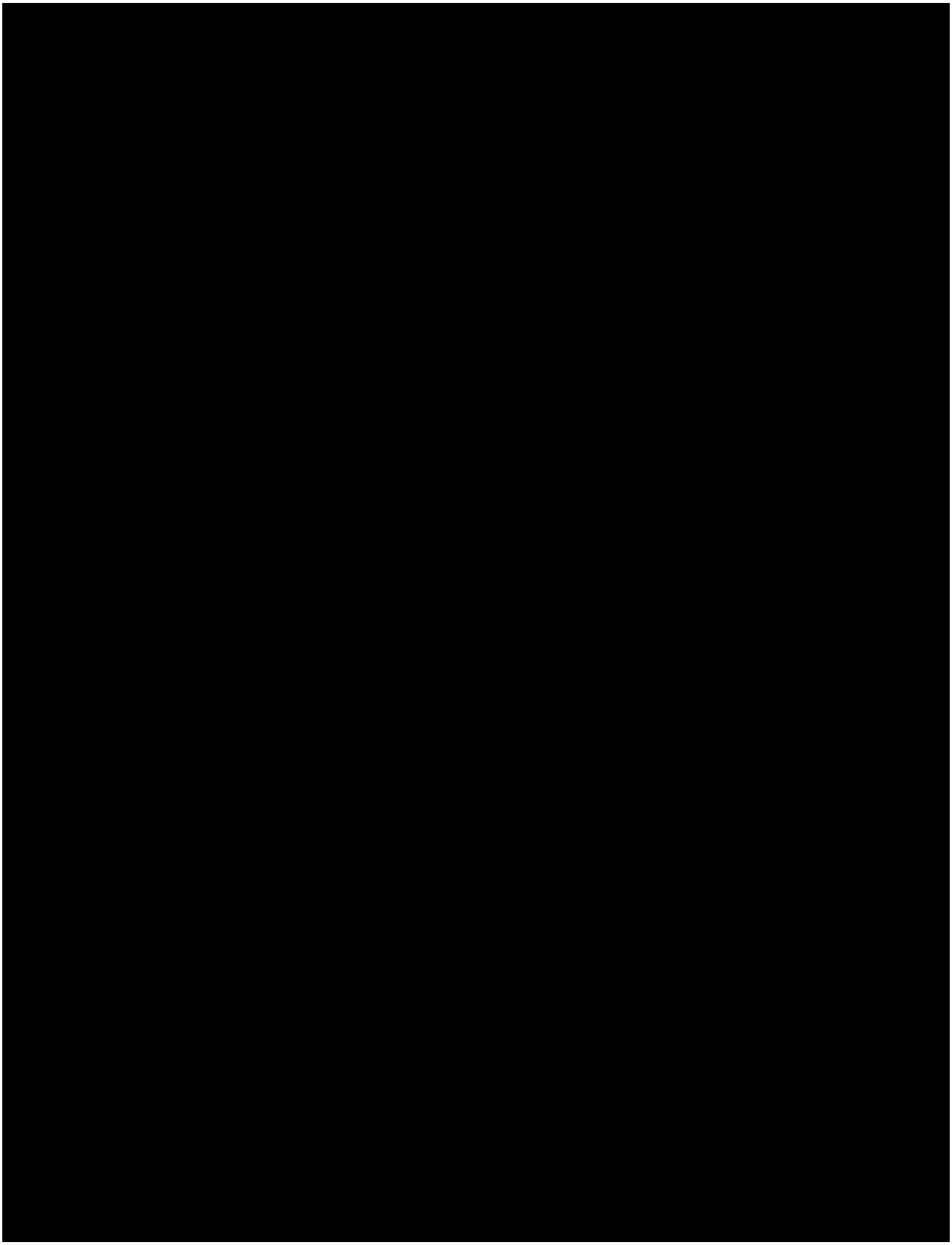
WHO (2013) Towards universal eye health: a global action plan 2014–2019 (WHA66.4) Available at: http://apps.who.int/gb/ebwha/pdf_files/WHA66-REC1/A66_REC1-en.pdf?ua=1 [Accessed 08/01/20]

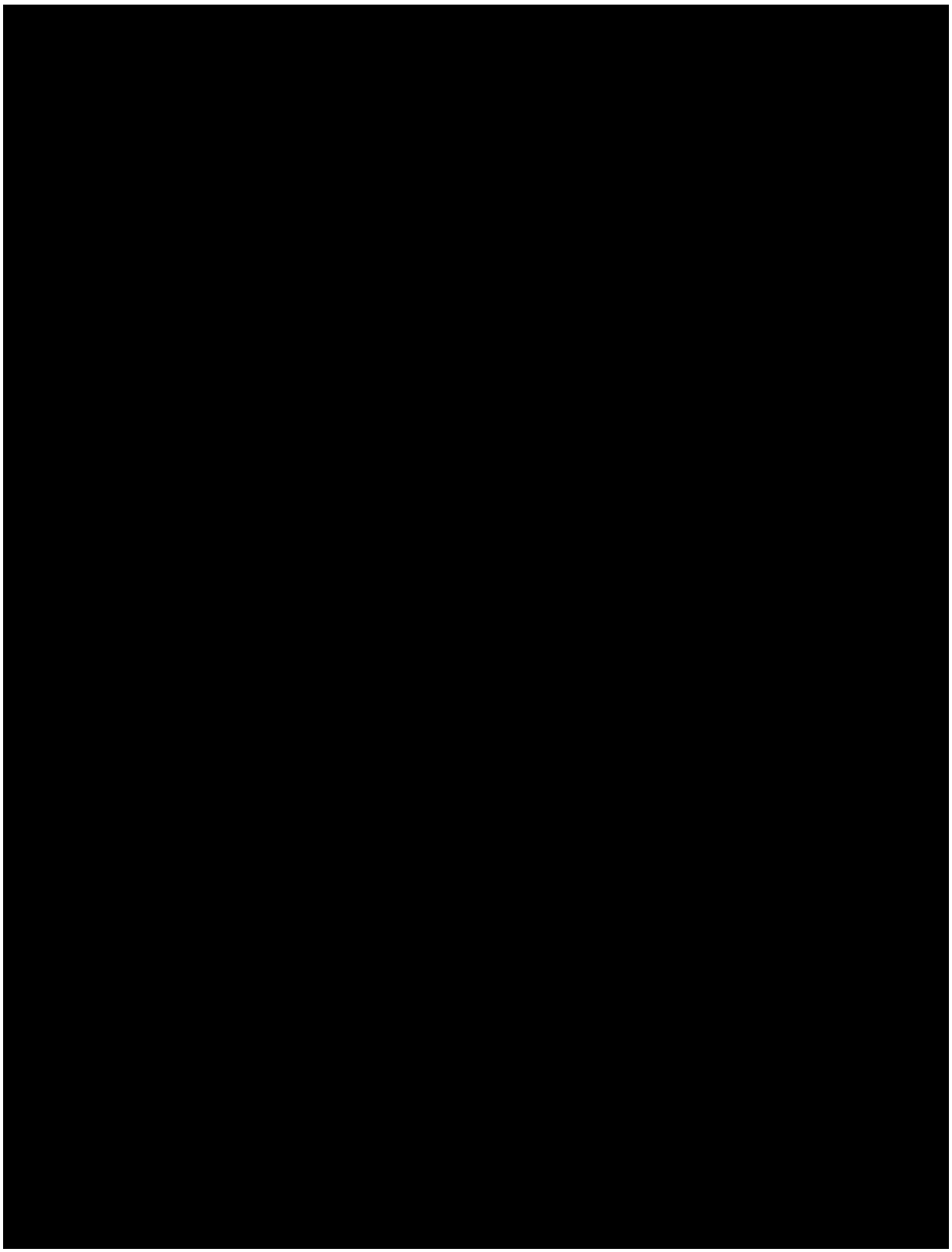
4.7. Published Article 2015

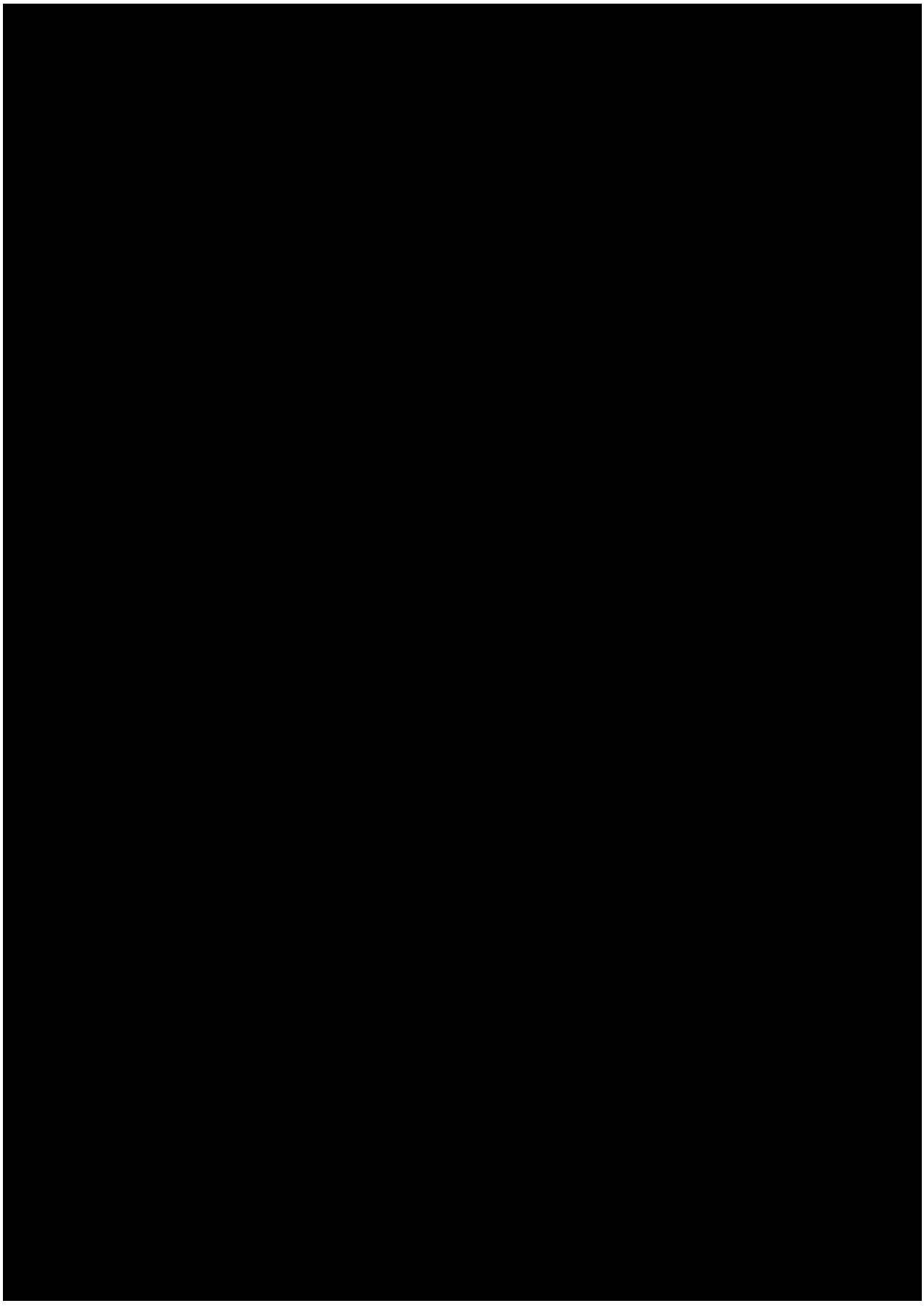


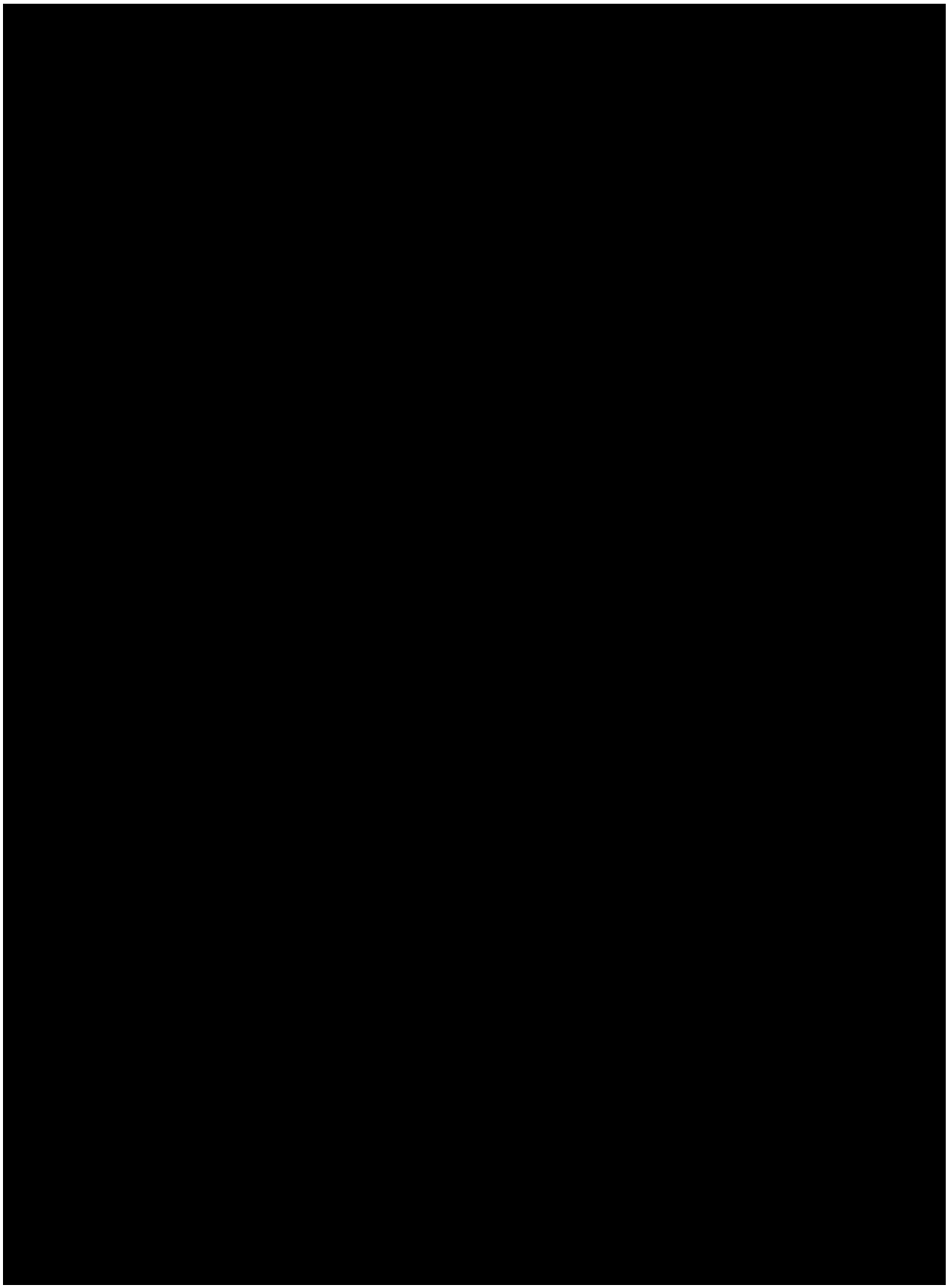


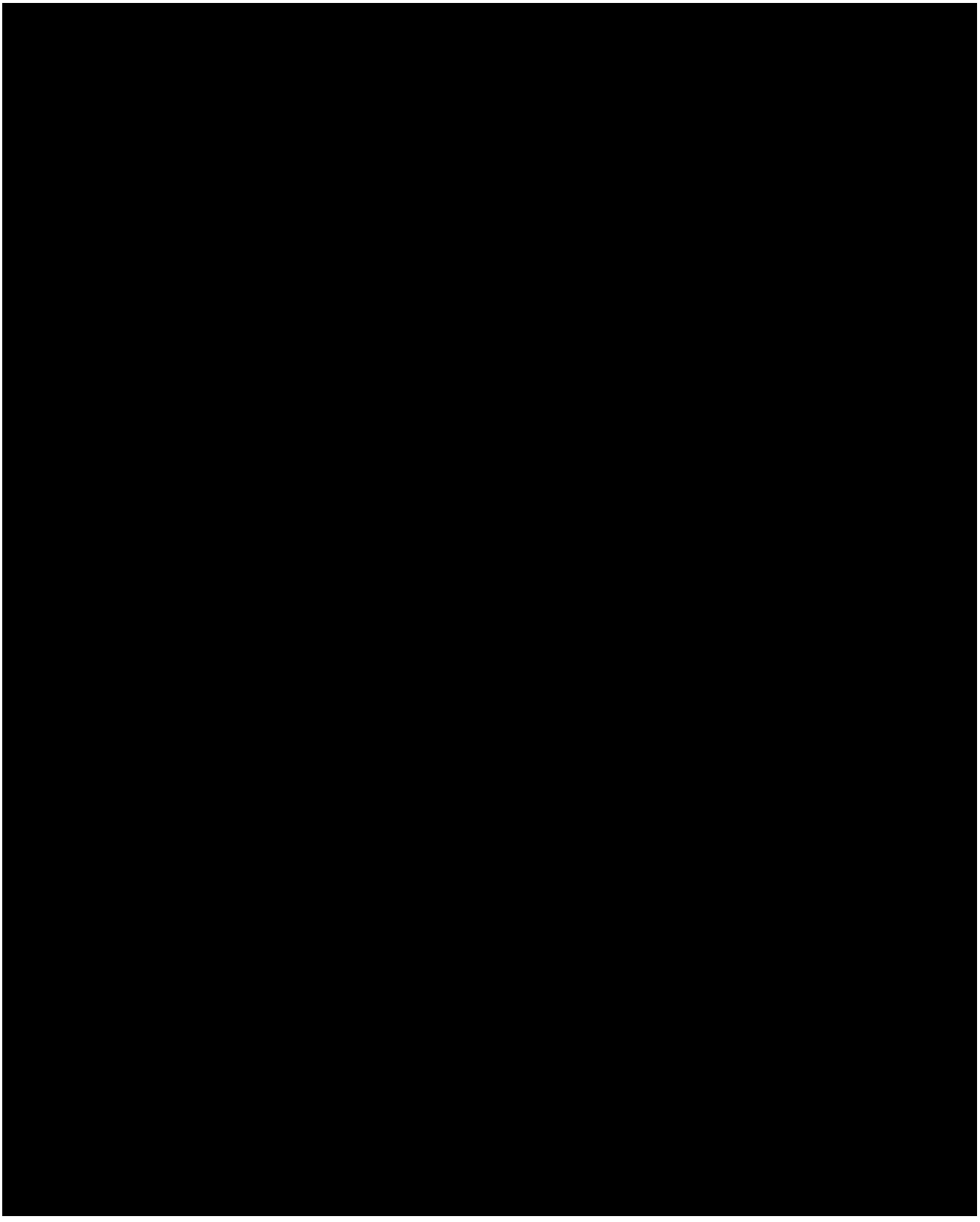




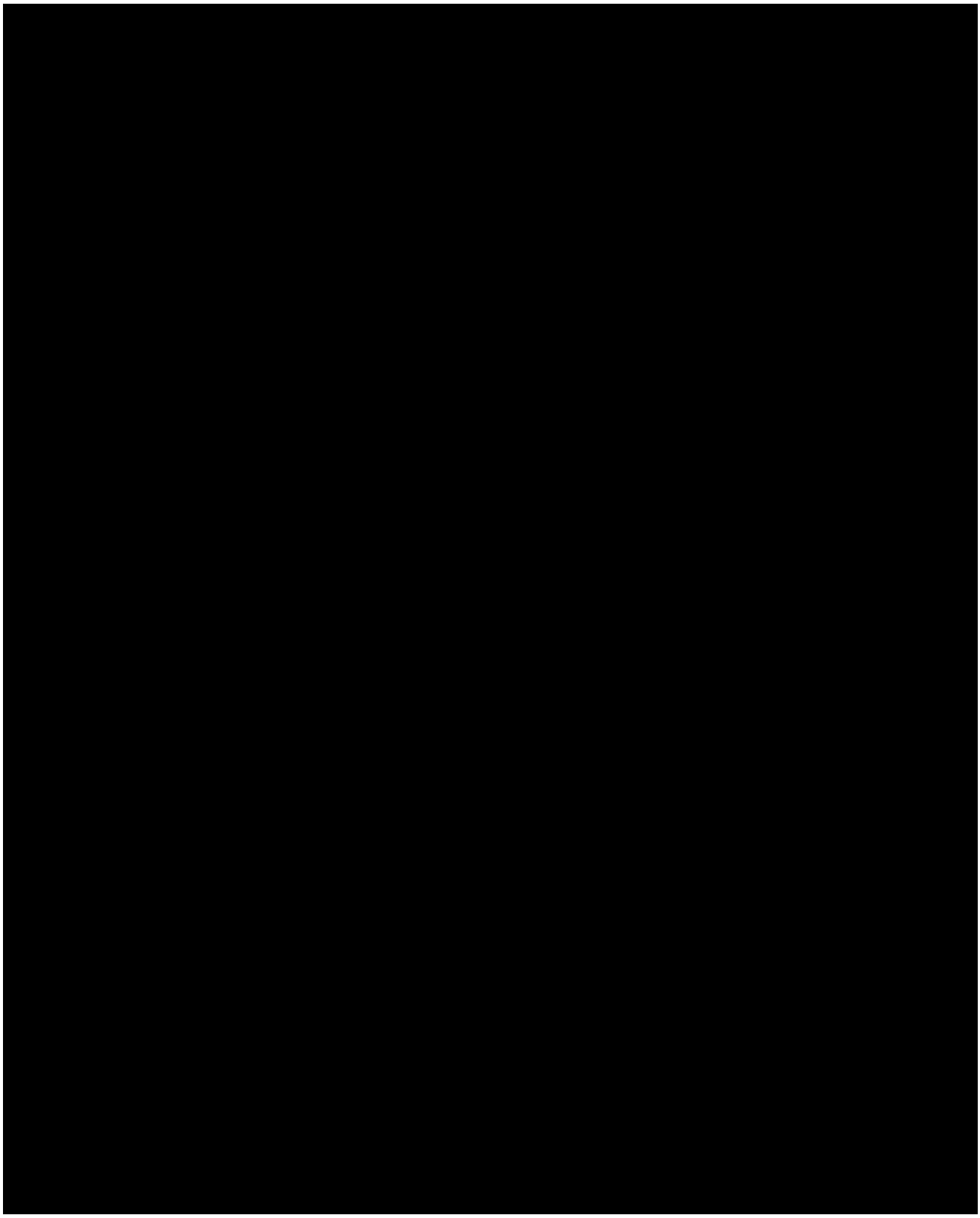


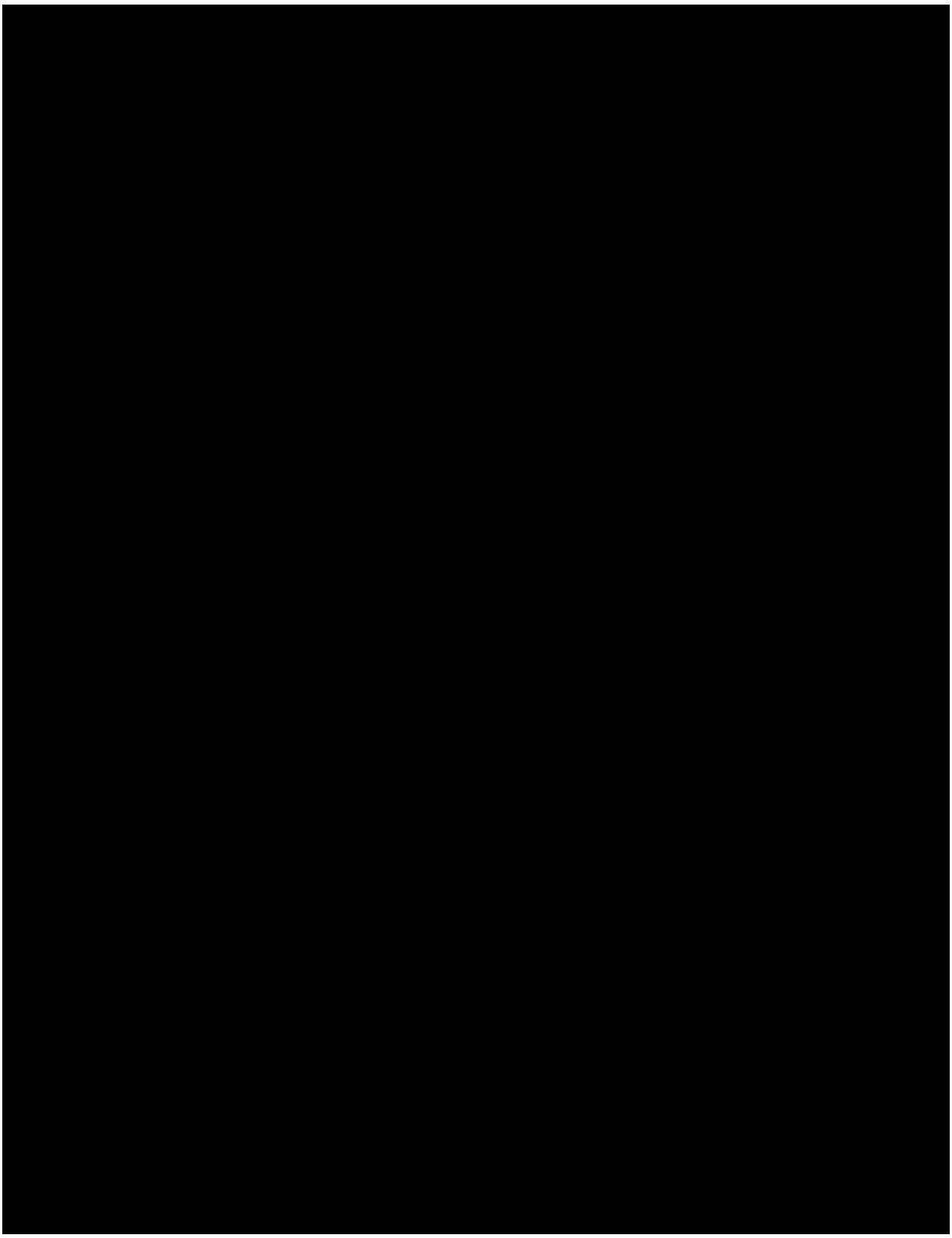


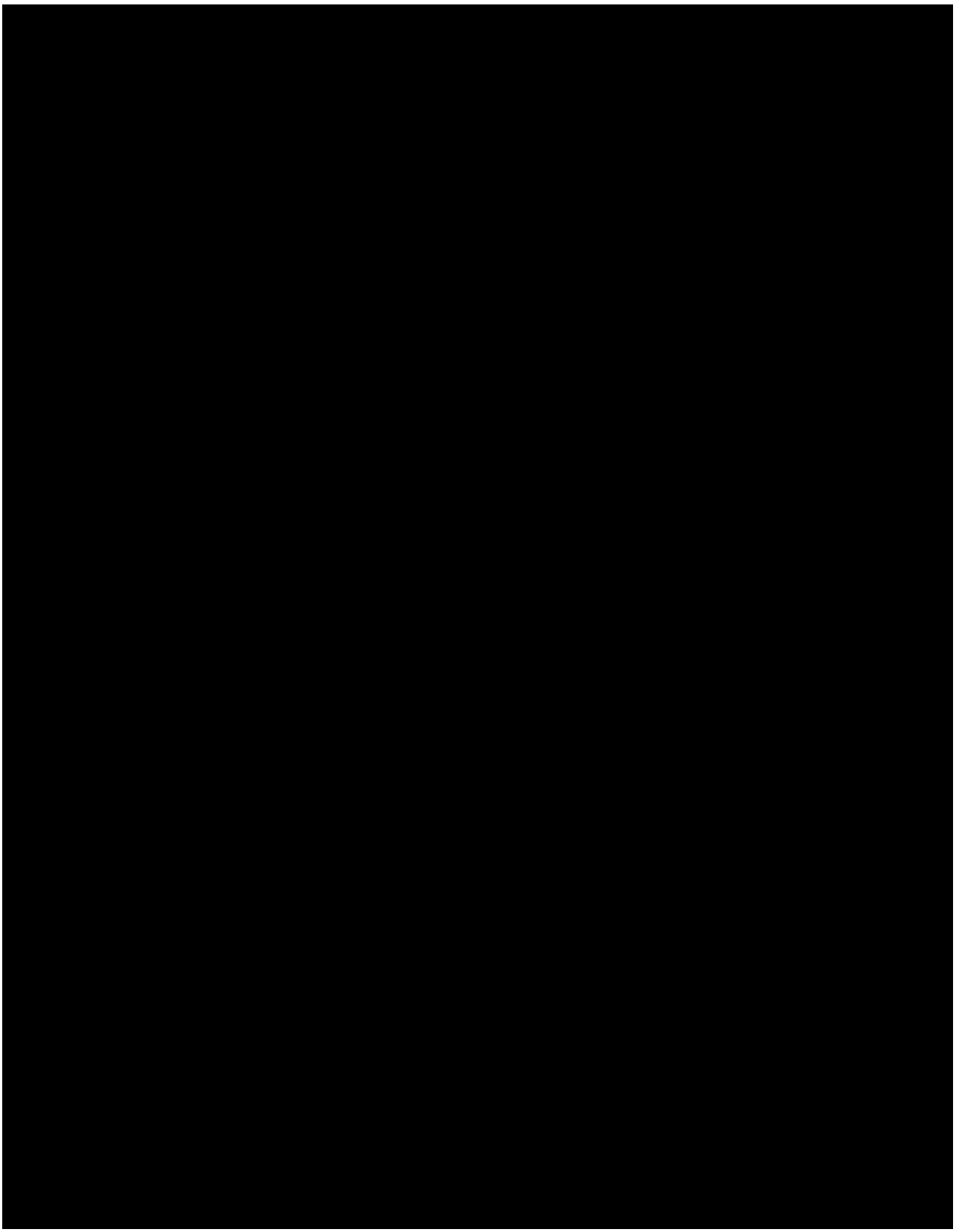


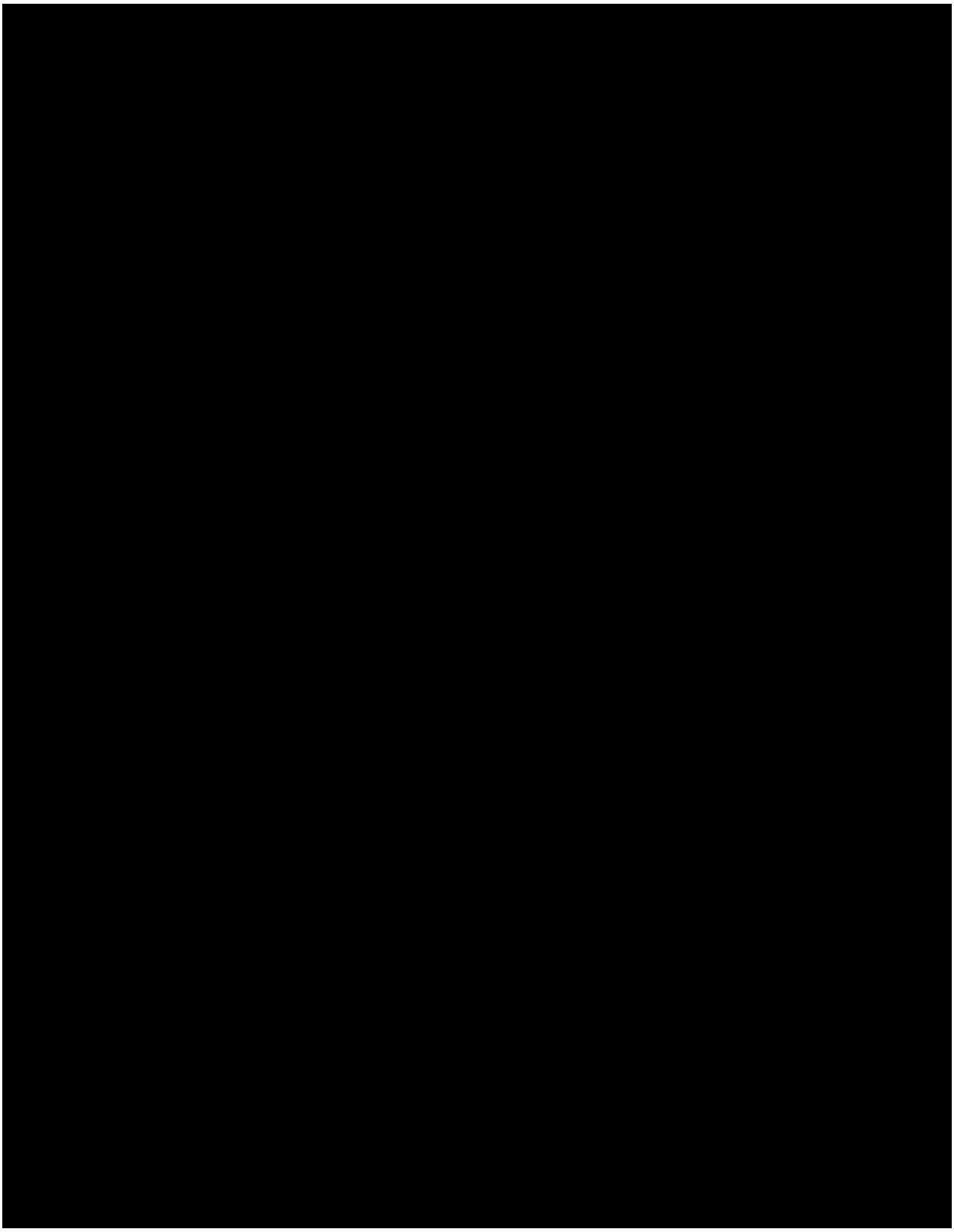


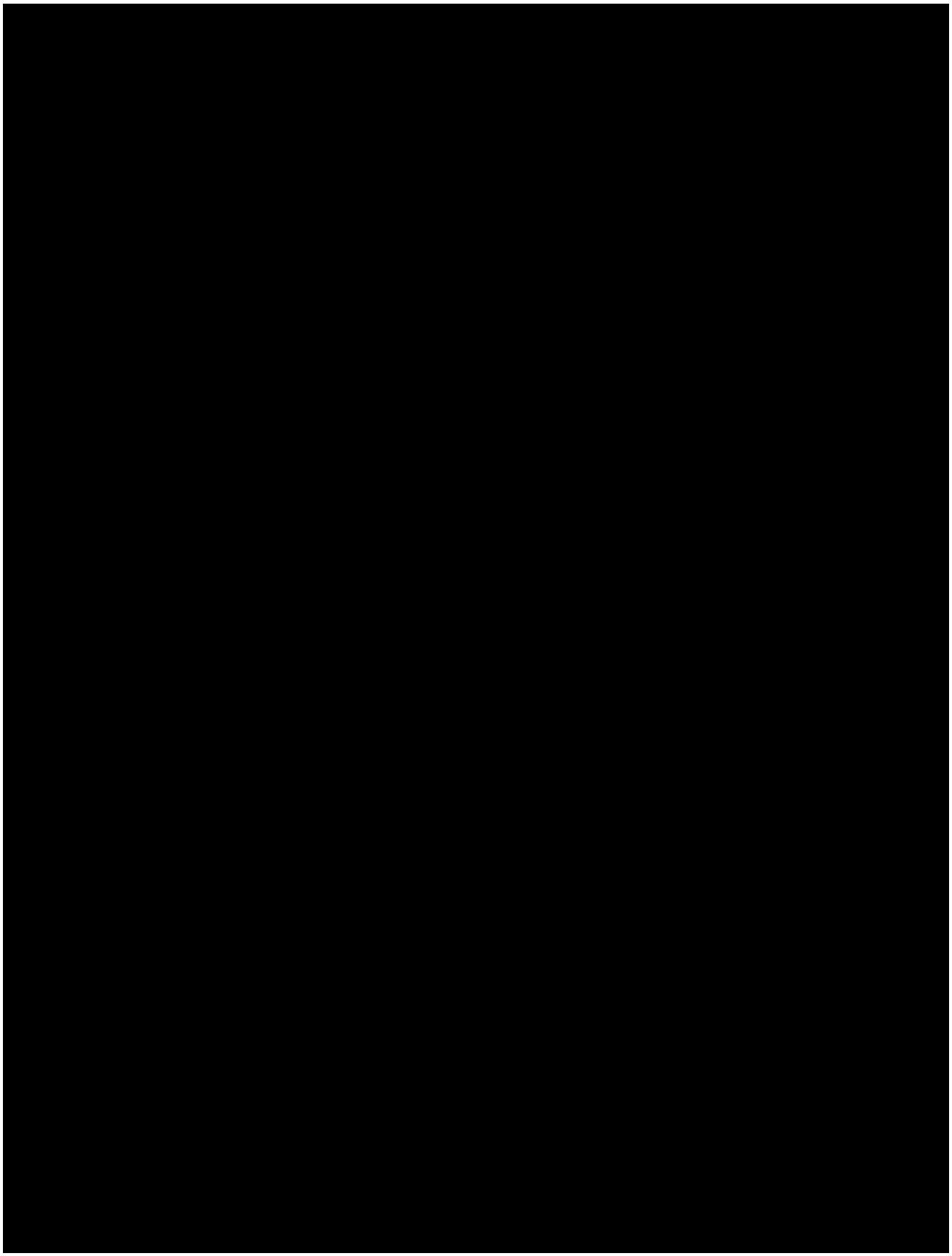
4.8 Published Article 2019

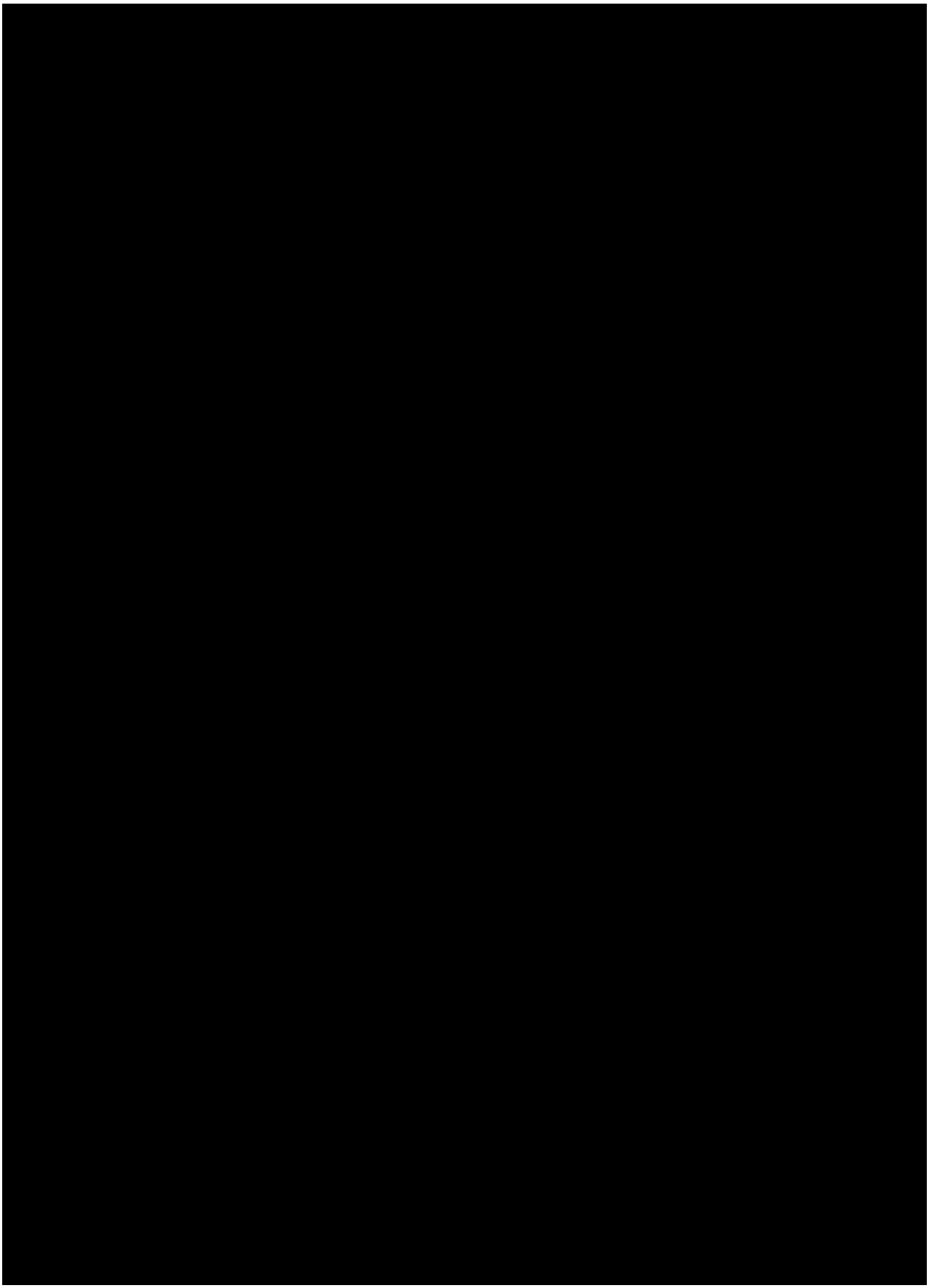


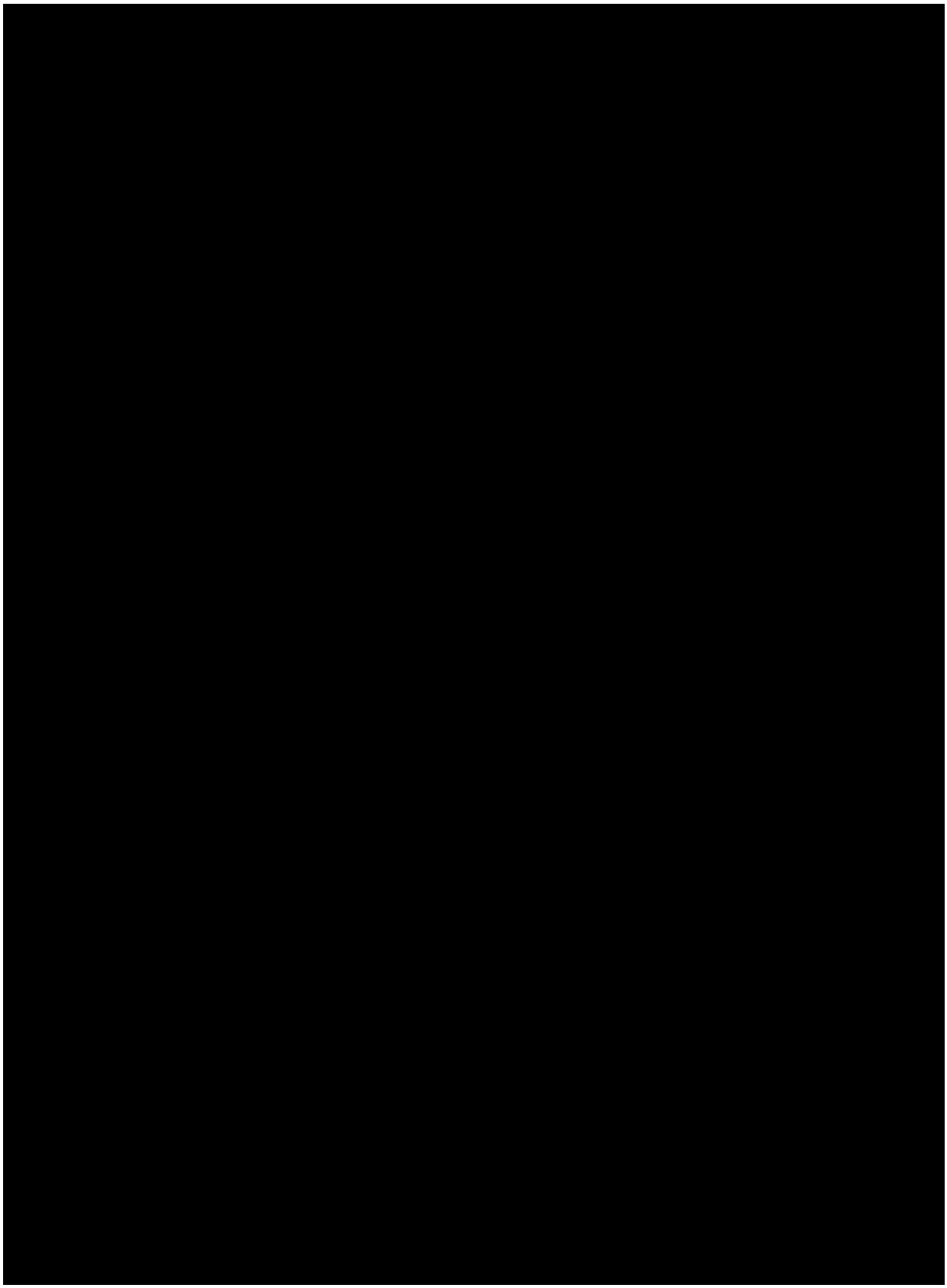


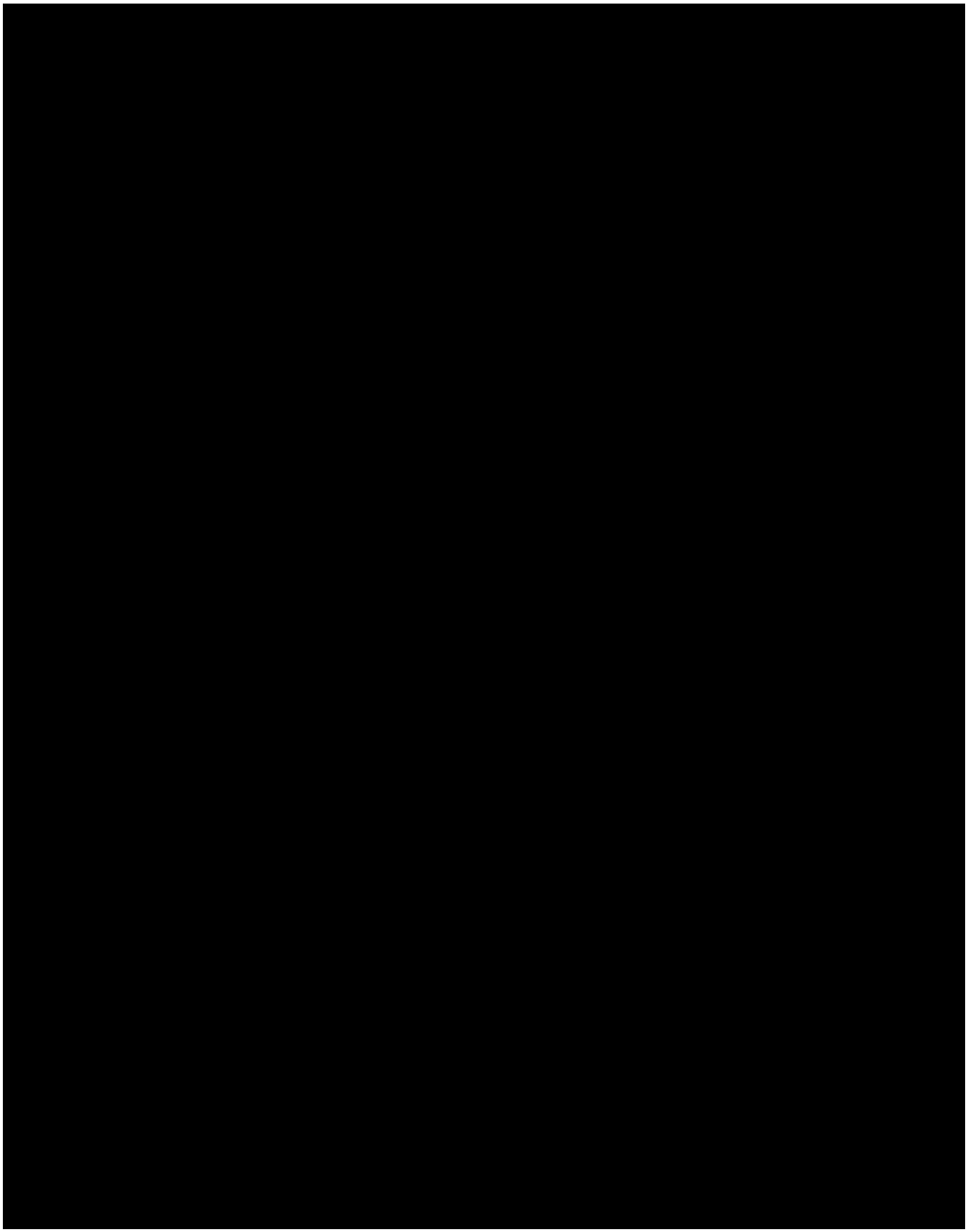












Chapter 5.

Summary of results and directions for future work

5.1. Summary

The work presented in this thesis describes a programme of research that encompassed the use of diagnostic tests evaluating ocular structure or function for the detection of COAG and other sight-threatening eye conditions. The work was disseminated in 4 peer reviewed publications (Fidalgo et al 2015, Dabasia et al 2015; Fidalgo et al 2018; Fidalgo et al 2019).

Diagnostic accuracy studies evaluate the ability of one or more 'index tests' to correctly classify patients as having a particular target condition, which is defined by an appropriate 'reference standard'. In 2003, reporting standards for this type of study design were developed and widely disseminated (Bossuyt et al 2003). The purpose of the Standards for Reporting of Diagnostic Accuracy Studies (STARD) checklist was to allow authors, journal editors and peer-reviewers to ensure that all relevant information is included in diagnostic accuracy studies. In parallel, an evidence-based methodological quality assessment tool (QUADAS) was developed (Whiting et al 2003) to assist systematic reviewers of diagnostic accuracy studies to evaluate risk of bias of studies included in the review. **Chapter 2** describes the quality of reporting and overall methodological quality of diagnostic accuracy studies that used perimetry to detect functional vision loss in glaucoma. Additionally, we investigated the impact of the publication of the STARD reporting standards on the quality of reporting by comparing articles published before and after the development and dissemination of the STARD checklist in 2003.

STARD compliance was poor with only 50% of the items adequately reported. Of all included articles in the study only 3 reported the use of the STARD checklist in the development of the paper. Less than 20% of journals recommend the use of STARD when reporting diagnostic accuracy studies compared to 50% of the same journals advising the use of CONSORT for the reports of RCTs. It is unclear whether the poor adoption of STARD by ophthalmology journals is the primary reason for the incomplete reporting of diagnostic accuracy studies identified in this study, however the requirement for authors to complete the STARD checklist prior to manuscript submission should be promoted.

Chapter 3 describes the development and diagnostic performance of a novel screening test that could potentially be used in conjunction with other clinical tests to detect COAG in a primary care setting. The Accelerator 4-Alternative Forced-Choice Flicker Test prototype (A4FTp) measures temporally-modulated flicker thresholds in regions of the visual field with high susceptibility to glaucomatous loss. We initially evaluated the psychometric properties of the A4FTp in 20 normal subjects who were tested multiple times over a period of 3 months. In addition, 4 randomly selected subjects underwent a total of 10 repetitions to study test-retest repeatability and learning effects. We showed that the A4FTp threshold algorithm with the shorter staircase termination criterion (T8) enabled rapid determination of flicker sensitivity in susceptible regions of the visual fields. Thresholds were repeatable and did not show any statistically significant learning effect over multiple repetitions.

To determine the diagnostic accuracy of the A4FTp, we compared the ability of the test to identify patients with COAG from a sample of 78 participants that included 40 subjects with COAG and 38 normal controls. The performance of the A4FTp was compared with Frequency Doubling Technology (FDT) perimetry (C20-5 programme) and iVue Spectral Domain Optical Coherence Tomography (SD-OCT). The accuracy of each test was determined by analysis of the area under the receiver operator characteristic curve (AUROC).

We found that test accuracy for the A4FTp was comparable with the FDT and SD-OCT for the detection of COAG. The results of this study demonstrated that with further refinement, the A4FTp could potentially have a future role in glaucoma detection.

Visual impairment disproportionately affects the elderly due to the increased risk of sight-threatening eye disease with age. Given that a high proportion of sight loss is preventable, there is a compelling case for early detection and referral for timely therapeutic intervention. Previous studies have found that no single test has sufficient predictive power to detect sight-threatening eye disease.

Chapter 4 describes the evaluation of the diagnostic performance of a number of visual function and structural tests for the detection of COAG and other sight-threatening eye disease in a representative sample of elderly subjects. Five hundred and five subjects underwent 4 index tests conducted by the author, who was unaware of subjects' ocular status. FDT and MMDT were used in supra-threshold mode. iVue SD-OCT measured GGC and RNFL thickness. The

diagnostic reference standard was full ophthalmic examination by an experienced clinician who was masked to index test results. The SD-OCT was the most effective in identifying subjects with COAG. Our results showed we would have identified 25 out of the 26 subjects diagnosed with definite COAG. The inferior RNFL thickness showed the greatest diagnostic accuracy, with a sensitivity of 77% and specificity of 95%. In a secondary analysis of the data, we also established that the combination of reduced visual acuity (VA <6/12), abnormal FDT, and peripapillary RNFL thickness outside 99% normal limits had the best overall discriminatory power for the detection of any sight-threatening eye disease, with a sensitivity estimate of 61% and specificity 79%, with similarly high positive and negative predictive values. The results provide useful data to inform the development of larger, multicentre population studies to evaluate the effectiveness and cost-effectiveness of screening for eye disease in the elderly.

5.2 Directions for future work

Since completing the work on Chapter 2 updated versions of STARD and QUADAS have been published (Whiting et al 2011; Bossuyt et al 2015). Consequently, future updates of the systematic review should utilise these updated standards. It is unclear whether simply highlighting poor reporting will lead to continuing improvements in methodological quality and standards of reporting of studies of diagnostic test accuracy. It may require lobbying of journal editors to update their guidance to authors to include reference to STARD. Most journals currently require adherence to CONSORT guidelines for the publication of RCTs.

The work described in Chapter 3 showed the potential of a new screening test for COAG, detection. This proof of concept study demonstrated that moderate to advanced COAG could be effectively detected using a small number of strategically placed flicker stimuli. Further development of the A4FTp is likely to involve optimisation of stimulus size number and location. This would be followed by the development of an appropriate normative database for the perimeter. Although the algorithm was able to quickly determine flicker thresholds at the four test locations, the use of a supra-threshold strategy would further reduce test times. A supra-threshold strategy may be more appropriate for case-finding in the general population. We are currently in the process of optimising the A4Ftp on a touch screen tablet display to increase its portability.

The results described in Chapter 4 showed the good predictive power of a small battery of screening tests to identify sight-threatening eye disease. Although we were able to recruit a relatively large sample of elderly subjects that were broadly representative of the population to be screened, the prevalence of the individual target conditions was low, which led to wide confidence intervals around the sensitivity and specificity estimates of test performance. Increasing the sample size would improve the precision around these estimates and also potentially provide greater ethnic diversity. Furthermore, the panel of screening tests identified in the set of subjects in the current study should also be independently verified on a separate validation sample to assess the generalisability of the findings. In parallel, a health economic analysis could be undertaken to investigate the cost-effectiveness of screening using the proposed model.

5.3. References

Ariyasu, R.G., Lee, P.P., Linton, K.P., LaBree, L.D., Azen, S.P., Siu, A.L., 1996. Sensitivity, Specificity, and Predictive Values of Screening Tests for Eye Conditions in a Clinic-based Population. Ophthalmology 103, 1751–1760.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Moher, D., Rennie, D., de Vet, H.C.W., Lijmer, J.G., Standards for Reporting of Diagnostic Accuracy, 2003. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Annals of Internal Medicine 138, W1-12.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C.W., Kressel, H.Y., Rifai, N., Golub, R.M., Altman, D.G., Hooft, L., Korevaar, D.A., Cohen, J.F., 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. British Medical Journal h5527.

Dabasia, P.L., Fidalgo, B.R., Edgar, D.F., Garway-Heath, D.F., Lawrenson, J.G., 2015. Diagnostic Accuracy of Technologies for Glaucoma Case-Finding in a Community Setting. Ophthalmology 122, 2407–2415.

Fidalgo, B.M.R., Crabb, D.P., Lawrenson, J.G., 2015. Methodology and reporting of diagnostic accuracy studies of automated perimetry in glaucoma: evaluation using a standardised approach. Ophthalmic and Physiological Optics 35, 315–323.

Fidalgo, B.R., Jindal, A., Tyler, C.W., Ctori, I., Lawrenson, J.G., 2018. Development and validation of a new glaucoma screening test using temporally modulated flicker. Ophthalmic and Physiological Optics 38, 617–628.

Fidalgo, B.R., Dabasia, P., Jindal, A., Edgar, D.F., Ctori, I., Peto, T., Lawrenson, J.G., 2019. Role of advanced technology in the detection of sight-threatening eye disease in a UK community setting. BMJ Open Ophthalmology 4, e000347.

Jindal, A., Ctori, I., Fidalgo, B., Dabasia, P., Balaskas, K., Lawrenson, J.G., 2019. Impact of optical coherence tomography on diagnostic decision-making by UK community optometrists: a clinical vignette study. Ophthalmic and Physiological Optics 39, 205–215.

Whiting, P., Rutjes, A.W., Reitsma, J.B., Bossuyt, P.M., Kleijnen, J., 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. Biomed Central Medical Research Methodology 3.

Whiting, P.F., 2011. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Annals of Internal Medicine 155, 529.

Appendix 1.

Summary table of studies included in quantitative analysis

First Author, Year	Sample size	Index Test	Reference Test
Katz 1993	5341	Humphrey full field 120	Optic disc examination, IOP and SAP
Mutlukan 1993	431	Multi-fixation campimeter	Aulhorn-Karmeyer classification
Wishart, 1993	56	OKP	Optic disc examination and SAP
Adachi, 1994	855	Noise field test	Optic disc examination and SAP
Christoffersen 1995	185	OKP	Optic disc examination and High-pass Resolution Perimetry
Sponsel 1995	Phase 1-143 Phase 2-176 Phase 3-1353	Henson Damato	Optic disc examination, IOP and SAP
Graham 1996	86	SWAP HRP Motion detection	Optic disc examination, IOP and SAP
Bosworth 1998	105	Motion automated perimetry (MAP)	Optic disc examination, IOP and SAP
Huang 1998	148	Dicon	SAP
Sim 1999	479	SAP	Optic disc examination, IOP and gonioscopy
Yamada 1999	240	FDT Damato	Optic disc examination, IOP and SAP
Burnstein 2000	29	FDT	SAP

Cello 2000	484	FDT	Optic disc examination, IOP and SAP
Fabre 2000	48	TOP FDT	Optic disc examination, IOP and SAP
Racette 2000	212	SWAP	Optic disc examination, IOP and SAP
Trible 2000	197	FDT screening FDT Full Treshhold	Optic disc examination and IOP
Vitale 2000	249	Dicon	Optic disc examination and gonioscopy
Paczka 2001	253	FDT	Optic disc examination, IOP and SAP
Bayer 2002	72	SWAP FDT	Optic disc examination and IOP
Horn 2002	639	FDT	Optic disc examination, gonioscopy and IOP
Iwasaki, 2002	14814	FDT	Optic disc examination and SAP
Wadood 2002	98	TOP FDT SITA-Fast	Optic disc examination, gonioscopy and IOP
Horn, 2003	307	FDT	Optic disc examination, gonioscopy, IOP and SAP
Detry-Morel 2004	1620	FDT	Optic disc examination and IOP
Babalola 2005	298	Motion sensitivity	Optic disc examination and IOP
Brusini 2005	123	Rarebit	Optic disc examination, gonioscopy and IOP
Fogagnolo 2005	80	FDT C-20 FDT N-30	General ophthalmic evaluation and SAP
Heeg 2005	452	FDT	SAP

Mansberger 2005	93	FDT	Optic disc examination and IOP
Robin 2005	704	FDT	Optic disc examination, IOP and SAP
Spry 2005	222	SWAP TMP FDT DAP RAP	Optic disc examination and IOP
Brusini 2006	318	FDT N-30 Matrix 30-2	Retinal nerve fiber layer (RNFL) and Optic disc examination
Gardiner 2006	218	FDT	Optic disc examination and SAP
Matsumoto 2006	135	Flicker FDT SAP	Retinal nerve fiber layer (RNFL), Optic disc examination, IOP and OCT
North 2006	100	FDT	Optic disc examination, IOP, gonioscopy and SAP
Pierre-Filho 2006	117	TOP FDT SITA Fast SITA Standard	Optic disc examination, gonioscopy and IOP
Sample 2006	246	SAP SWAP FDT HPRP	Optic disc examination and IOP
Ferreras 2007	294	SWAP FDT	Optic disc examination, IOP and SAP
Fortune 2007	185	SAP	Optic disc examination and output of the Moorfields regression analysis from the Heidelberg Retina Tomograph

Horn 2007	109	FDT SWAP	Optic disc examination, IOP, gonioscopy, SAP and papillometry
Iwase 2007	2892	FDT	Optic disc examination, IOP, gonioscopy and SAP
Kumar 2007	399	FDT	Conventional hospital-based method
Leeprechanon 2007a	77	FDT SWAP	Ophthalmological examination
Leeprechanon 2007b	92	FDT GHT FDT PSD	Optic disc examination, gonioscopy and IOP
Spry 2007	53	FDT 24-2-5 FDT N-30-5	Optic disc examination, gonioscopy and IOP
Wang 2007	4349	FDT	Optic disc examination
Cook 2009	105	FDT	Ophthalmological examination
Salim 2009	70	FDT	IOP, optic disc examination
Tafreshi 2009	338	SAP SWAP FDT	Optic disc examination, gonioscopy and IOP
Toth 2009	181	Matrix FDT	Optic disc examination, IOP, gonioscopy and SAP
Rowe 2010	197	Damato	Optic disc examination and SAP
Salvetat, 2010	108	FDT RBP PP	Optic disc examination and SAP
Zhong, 2010	160	SWAP	Optic disc examination, gonioscopy and IOP
Francis 2011	6082	FDT	Optic disc examination, gonioscopy and IOP
Gonzalez 2011	328	Pulsar FDT	Optic disc examination, SAP

Kamdeu, 2011	550	FDT	Optic disc examination, gonioscopy and IOP
Anton 2012	186	ATD Multichannel Functional Test	SAP, optic disc examination, optical coherence tomography (OCT) and IOP
Horn 2012	588	FDT Pre-perimetric FDT Perimetric	Optic disc examination, SAP
Horn 2014	171	FDF	SAP, optic disc examination, OCT, gonioscopy and IOP
Kanadani 2014	95	FDT	Optic disc examination, IOP, gonioscopy and SAP
Prokosch 2014	91	FDT FDF SAP	OCT
Dabasia 2015	505	FDT MMDT ORA OCT	Optic disc examination, IOP, gonioscopy, fundus photography and SAP
Ghazali 2015	518	SAP	Optic disc examination and IOP
Matsumoto 2015	159	Clock Chart	Optic disc examination, IOP and SAP
Rosen 2015	130	PERCEPT	Optic disc examination, IOP, gonioscopy and SAP
Boland et al, 2016	6797	FDT	Fundus photography
Mwanza 2016	224	FDT	Optic disc examination, IOP and SAP
Olsen 2016	97	DMCO	Optic disc examination and SAP
Johnson et al, 2017	206	VFE	Optic disc examination, IOP and gonioscopy

Takahashi 2017	141	G-Dynamic and GST	Optic disc examination and SAP
Kita 2017	18	Circumpapillary microperimetry (MP)	Optic disc examination, IOP and gonioscopy
Meethal 2017	104	Eye Movement Perimetry (EMP)	Optic disc examination and SAP
Abreu-Gonzalez 2018	202	OCULUS Smartfield perimeter (SPARK strategy)	Optic disc examination and IOP
Fidalgo 2018	78	A4FTp, FDT	Optic disc examination, IOP, gonioscopy and SAP
Olsen 2018	627	DMCO	Optic disc examination, IOP, gonioscopy and SAP
Meethal 2019	104	Eye Movement Perimetry (EMP)	Optic disc examination and SAP

Abbreviations: A4FTp - Accelerator 4-Alternative Forced-Choice Flicker Test prototype; DAP - Detection acuity perimetry; DMCO - Damato multifixation campimetry; EMP - Eye Movement Perimetry; FDF – Flicker defined form; FDT – Frequency doubling technology; HRP - High-pass resolution perimetry; MMDT – Moorfields motion displacement test; MP – microperimetry; OCT - Optical coherence tomography; ORA – Ocular response analyser; OKP - Oculo-kinetic perimetry; PP - Pulsar perimetry; RAP - Resolution acuity perimetry; RBP – Rarebit perimetry; SAP – Standard automated perimetry; SWAP - Short Wavelength Automated Perimetry; TMP - Temporal modulation perimetry; VFE - Visual field extent

Appendix 2.

A4FTp flicker algorithm

```
function FlickerScreen(f,n)
% Two-screen version of a 4-patch peripheral flicker testing program
% Control screen should not be visible to test subject
% After first run, set windows to be non-overlapping on control screen
% To run, choose 'Screening' or 'Choose Frequency' at the prompt
% Screening starts the flicker at a moderate level suitable for patients
% Choose frequency starts the flicker close to normal thresholds
% Fixate the central dot and press any key to start
% Press the 1, 2, 4 or 5 key on the number pad to match the flicker location
% If no flicker seen, press the same key as for the previous trial
% If unsure, press 0 to repeat same levels (but different location)
% Press q + RETURN to quit
% If program crashes, type Screen('CloseAll') and rerun program
% For City system, change line 16 to PsychToolBox call, and activate line 42
% Figure 1 shows the actual waveform presented on each trial (black trace)
% and at 10x scale for visibility of the waveform at low amplitudes (green trace)
% The results are for upper left, upper right, lower left and lower right
% Results should be accurate to about +/- 1 decilog
% Ignore the text printing out after the results

global V
global PCS
clear all

% Choose screen with maximum id - the secondary display:
%screens=Screen('Screens')
%screenid = max(Screen('Screens'));
%screenid=Max(Screens);
screenid = 0; %Change value to 2 to force presentation on second (144 Hz)
screen

% Open a fullscreen onscreen window on that display, choose a background
% color of 128 = gray with 50% max intensity:
%win = Screen('OpenWindow', screenid, 128, rect, [], [], [], [], imagingMode);

cumpc=[1 0 0 0]

if isempty(which('Screen'))
        %configureTylerlabPTB %Calling PsychToolBox: specific to machine
        addpath('C:\Matlab\CTC\Psychtoolbox\PsychBasic\',0)
%C:\Matlab\CTC\Psychtoolbox\PsychBasic\PsychToolBox
end

if nargin < 1
        f = [];
end

if isempty(f)
        f = 25;
end
```

```
if nargin < 2
	n = [];
end

if isempty(n)
	n = 10;
end

% Close previous figure plots:
close all;

% Make sure this is running on OpenGL Psychtoolbox:
AssertOpenGL;
% maxcount=60; %Screen refresh rate
maxcount=120; %Screen refresh rate
disp('Screen refresh rate set to '), disp(maxcount)

% ShowHideWinTaskbarMex(0)
HideCursor
% load RESULTS
load Patients
dn=numel(Patients);
fpmi=input(['Patient Number (less than ' num2str(dn+1) ') or - 0 for automatic
number '],'s');
fpmi=str2num(fpmi);

% while numel(details1)>FlickerPatientNumberinput
%     FlickerPatientNumberinput=input('Patient Number does not exist - 0 for
automatic number ','s');
% end

if fpmi==0
	FlickerPatientNumber=numel(Patients)+1;
	name=input('Name ','s');
	dob=input('DoB (DD/MM/YYYY) ','s');
	gender=input('Gender (M/F) ','s')
	Patients{FlickerPatientNumber}.dob=dob;
	Patients{FlickerPatientNumber}.gender=gender;
	Patients{FlickerPatientNumber}.name=name;

Patients{FlickerPatientNumber}.FlickerPatientNumber=FlickerPatientNumber
;
	save Patients Patients
else
FlickerPatientNumber=fpmi;
Patients{FlickerPatientNumber}.name
end

viewingeye=input('Eye (L/R) ','s');
Eyefield=1; %setting right field option
if viewingeye=='R'
	Eyefield=-1;
elseif viewingeye=='r'
	Eyefield=-1;
```

```
end

f=60 %Temporal frequency
% f=input('Flicker frequency (5/10/15/30/60)? ','s')
ff=input('Flicker frequency 30 or 60 (0/1))? ','s');
if ff=='0'
        f=30
end

% Initial stimulus params for the stimulus patches:
res = 1*[323 323];
sc = 300; %space constant
border=50;%width of black border
inc=0.27; %staircase increment - 0.27 is 0.1 log unit
%inc=0; %staircase increment - 0.27 is 0.1 log unit
dec=inc/2; %staircase decrement
NumTrials=120; %maximum number of trials
n=8; %number averaged
gamma=1.9;
pcInit=30; %Initial percent contrast


% Disable synctests for this quick demo:
oldSyncLevel = Screen('Preference', 'SkipSyncTests', 0);

tw = 1200;
th = 900;
% x=tw/2;
% y=th/2; %Centre of screen
x=tw/3;  y=th/3;

white=WhiteIndex(0); black=BlackIndex(0);
%Using gray-0.5 as the mean level puts the mean halfway between two steps
%Butit cannot match the steady level, so cannot be used
%we will have to rely on using high frequencies with threshold > 1 step.

% bgLum = 0.25;
% gray=floor(bgLum*white+(1-bgLum)*black);
gray=floor((white+black)/2);

[win, wRect]=Screen(screenid,'OpenWindow',[gray,0,0])
vbl_flip_int=Screen('GetFlipInterval',win)
expectedtime=vbl_flip_int*maxcount

Screen('FrameOval',win,0,OffsetRect(CenterRect([1 1 40 40],wRect),-
Eyefield*120,0));
Screen('FrameOval',win,[0,0,0],OffsetRect(CenterRect([1 1 sc sc],wRect),-
Eyefield*x/2-Eyefield*120,y),4,4); %CWT
Screen('FrameOval',win,[0,0,0],OffsetRect(CenterRect([1 1 sc sc],wRect),-
Eyefield*x/2-Eyefield*120,-y),4,4); %CWT
Screen('FrameOval',win,[0,0,0],OffsetRect(CenterRect([1 1 sc
sc],wRect),Eyefield*x-Eyefield*120,y/2),4,4); %CWT
Screen('FrameOval',win,[0,0,0],OffsetRect(CenterRect([1 1 sc
sc],wRect),Eyefield*x-Eyefield*120,-y/2),4,4); %CWT
```

```
% Perform initial flip to gray background and sync us to the retrace:
vbl = Screen('Flip', win,0);
disp('Press any key to start')
pause
ts = vbl;
tic; %test time start
fig = figure('Position',[100 650 650 420]);

%Psychophysics loop: Run to NumTrials:
pc = pcInit*ones(1,4);
cumpc=pc;
PCS={pc(1) pc(1) pc(1) pc(1)};
V=ones(1,4);

if Eyefield==1
        locKeys = '1739';
else
        locKeys = '3917';
end

locKeyIndex = zeros(1,numel(locKeys));
for i = 1:numel(locKeys)
        locKeyIndex(i) = KbName(locKeys(i));
end

ListenChar(2)
while max(V)>0.15
        framecount = 0;
        locSwitch=round(0.5+rand*4);
%    locSwitch=1;
        switch locSwitch %CWT
                case 1
                        loc = [-Eyefield*x/2-Eyefield*120,y];
                case 2
                        loc = [-Eyefield*x/2-Eyefield*120,-y];
                case 3
                        loc = [Eyefield*x-Eyefield*120,y/2];
                case 4
                        loc = [Eyefield*x-Eyefield*120,-y/2];
        end %switch

prioritylevel=MaxPriority(win);
Priority(prioritylevel);

% tic %stym duration (Note this tick will overwrite the tic for the test
% duration)
        % Animation loop:
        while framecount < maxcount*0.75
                framecount = framecount + 1;
                %Gabor temporal envelope around 'gray' with pc peak contrast
                tc(framecount)=gray+pc(locSwitch)*exp(-((framecount-
maxcount/3)^2/((maxcount/6)^2)))*cos(2*pi*f*framecount/maxcount);   %JAS:
Gabor temporal envelope with pc peak contrast
                % Gamma adjusted tc
```

```
        tcCal(framecount)=floor(tc(framecount)^(1/gamma)*(127/(127^(1/gam
ma))));
        Screen('FillOval',win,[tcCal(framecount),0,0],OffsetRect(CenterRect([
1 1 sc-8 sc-8],wRect),loc(1),loc(2))); %CWT
                Screen('FrameOval',win,0,OffsetRect(CenterRect([1    1    40
40],wRect),-Eyefield*120,0)); %Fixation spot
        Screen('Flip',win,0,0,0);
% WaitSecs(0.0144)
        end %while framecount
%  toc %part of tic for stym duration
        Beeper(400,1,0.1)
%vbl_time=vbl-ts
        Priority(0);
        commandwindow
        figure(fig),   plot((1:framecount)/maxcount,10*round(tc(1:framecount)-
gray)+gray,'g')
        hold on; plot((1:framecount)/maxcount,round(tc(1:framecount)),'k');
        axis([0 1 0 255]); drawnow
        title('Black is true waveform, green is x10')
        xlabel('Time (sec)')
        ylabel('Luminance levels')
        hold off

        while 1 %checking for keypress
                [KeyIsDown, endrt, KeyCode] = KbCheck;
                if KeyIsDown
                        break;
                end
        end %while KBCheck

        if KeyCode(locKeyIndex(locSwitch))
                pc(locSwitch)=max(1,pc(locSwitch)/(1+dec));
        elseif KeyCode(96)
                %do not change values
        elseif any(KeyCode(locKeyIndex))
                pc(locSwitch)=min(127,pc(locSwitch)*(1+inc));
        else
%               q=input('Do you want to quit (q) or continue(c)?','s')
                fprintf('Do you want to quit (q) or continue(c)?\n')
                [~,KeyCode(:)] = KbWait(-3,2);
                if KeyCode(KbName('q')) == 1
                        ListenChar(0)
                        Screen('CloseAll');
                        cumpc
                        return
                else
                        disp('CONTINUING')
                end
        end
        cumpc=[cumpc;pc];
        TrialNum=size(cumpc);
        PCS{locSwitch}=[PCS{locSwitch};pc(locSwitch)];
        locSwitch, sl=length(PCS{locSwitch}) %printing location, staircase
length
```

```
        if sl>n
                Amplitudes=[PCS{1}(end)  PCS{2}(end)  PCS{3}(end)
PCS{4}(end)]
                V(locSwitch)=std(log10(PCS{locSwitch}(sl-n+1:sl)));
        end

        if TrialNum(1)>NumTrials
                ListenChar(0)
                display('Run did not stabilize: Rerun')
                Screen('CloseAll')
        cumpc;
                return
        end %if

end % while max(V)>0.15
ListenChar(0)
toc;%test time end
testtime=toc;
% cumpc=[FlickerPatientNumber 0 0 0; cumpc]
% save cumpc

NumberOfTrials=length(cumpc(:,1));
Thresholds=[PCS{1}(end)  PCS{2}(end)  PCS{3}(end)  PCS{4}(end)]/1.27
%Converts lut steps to percentage
disp('(Minimum measurable threshold = 1)')
% Decilog_Sensitivities=floor(log10(100./Thresholds)*10)
disp('(Maximum measurable decilog sensitivity = 20 (= 2 log units) )')
result=[FlickerPatientNumber  Eyefield/2+1.5  f  NumberOfTrials  Thresholds
testtime];
result
c=clock;
time=100*floor(c(4))+floor(c(5));
fpmi
flickername=strcat(Patients{FlickerPatientNumber}.name,date,num2str(time),
'screening')
% RESULTS=[RESULTS;result];
save(flickername,'result','-ascii')
% whos
xlswrite(flickername,result)

%The following lines save into the txt file the demographics data for the
%patient being tested

outfile=fopen('ScreenDemographics.txt','at')

fprintf(outfile,[num2str(FlickerPatientNumber)]);
fprintf(outfile,',');
fprintf(outfile,date);
fprintf(outfile,',');
fprintf(outfile,Patients{FlickerPatientNumber}.name);
fprintf(outfile,',');
fprintf(outfile,Patients{FlickerPatientNumber}.gender);
fprintf(outfile,',');
fprintf(outfile,Patients{FlickerPatientNumber}.dob);
fprintf(outfile,'\n');
```

```
% ShowHideWinTaskbarMex(1)
% A final synced flip, so we can be sure all drawing is finished when we
% reach this point:
tend = Screen('Flip', win, 0);

ShowCursor

% pause
%KbWait;
% Close window, release all resources:
%Screen('CloseAll');
display('End of Run')
Screen('Close',win);
return;
```

Appendix 3.

Patient Acceptability Questionnaire

Whilst the pupils were dilating, subjects were asked to complete a questionnaire regarding the acceptability of each of the index tests, attached bellow.

CITY UNIVERSITY LONDON

Screening Study of Equipment and its Impact in Eye Care
Questionnaire of User Acceptability of Screening Tests

Date of Examination:…… Subject ID SEC ...……...

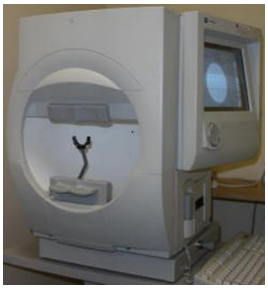
Unless otherwise stated, please fill one circle for each question using black or blue ink

For Questions 1 – 5, please indicate whether you agree or disagree with the statements relating to your views on the screening tests carried out on you today, using the nine-point scale provided.

EXAMPLE:

	Disagree						**Agree**
The screening test was uncomfortable	O	●	O	O	O	O	O

Question 1: Humphrey visual fields (Location: small room on Level 4) – responding to white flashes on a screen

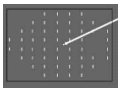


	Disagree						**Agree**
The screening test was uncomfortable	O	O	O	O	O	O	O
The test was too long	O	O	O	O	O	O	O
The test was difficult to undertake	O	O	O	O	O	O	O

Yellow-orange fixation light

Question 2: MMDT (Location: larger room on Level 6) – responding to moving white lines

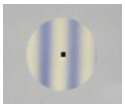
	Disagree				Agree		
The screening test was uncomfortable	O	O	O	O	O	O	O
The test was too long	O	O	O	O	O	O	O
The test was difficult to undertake	O	O	O	O	O	O	O



White fixation spot

Question 3: FDT (Location: larger room on Level 6) – responding to flickering white and black bars

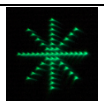
	Disagree				Agree		
The screening test was uncomfortable	O	O	O	O	O	O	O
The test was too long	O	O	O	O	O	O	O
The test was difficult to undertake	O	O	O	O	O	O	O

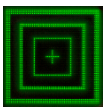


Black fixation square

Question 4: iVue OCT (Location: larger room on Level 6) – instrument captures images of the back of your eye

	Disagree				Agree		
The screening test was uncomfortable	O	O	O	O	O	O	O
The test was too long	O	O	O	O	O	O	O
The test was difficult to undertake	O	O	O	O	O	O	O





Green star or cross target

Question 5: ORA (Location: larger room on Level 6) – 'puff of air' in the eye to measure your eye pressure

	Disagree				Agree		
The screening test was uncomfortable	O	O	O	O	O	O	O
The test was too long	O	O	O	O	O	O	O
The test was difficult to undertake	O	O	O	O	O	O	O

Target of green spot within four red lights

Question 6 **If you have any further comments on the acceptability of tests undertaken today, or on any other aspect of the study, please write them in the box below:**

Appendix 4.

Other published articles

Fundus and OCT images collected as part of the work described in Chapter 4 was used to study the diagnostic decision-making of UK community optometrists. This generated a further peer-reviewed publication on which I was a co-author.