Keratoconus detection and personalised progression modelling using computational methods

Author: Howard Maile

Supervisors: Prof. Stephen Tuft, Dr. Nikolas Pontikos

A thesis submitted in fulfilment of the requirements for the degree of MPhil in Ophthalmology at University College London

UNIVERSITY COLLEGE LONDON

January 2023
Abstract

Keratoconus is a common corneal disorder in young adults characterised by bilateral thinning and distortion of the cornea that disrupts this refractive system leading to blurred vision. Untreated, it can progress to cause blindness and is a leading cause of visual loss in young adults globally. Early detection, understanding disease progression and administering optimal treatment are thus important areas of research. Initially, the existing literature on the algorithmic detection of subclinical keratoconus was surveyed and critically evaluated which resulted in a comprehensive systematic review as well as the identification of avenues for further research. This led to the exploration of using convolutional neural networks (CNNs) applied to raw Anterior Segment Optical Coherence Tomography (AS-OCT) images. The results showed that raw images can be used to classify disease severity (early versus late stage keratoconus) with validation accuracy of 96.5% and provide the groundwork for further research. Subsequent chapters deal with the problem of predicting and analysing keratoconus progression. A prognostic model was developed to predict keratoconus progression to requiring corneal cross-linking (CXL) with explained variation of 33%. Given a series of corneal parameters at the first appointment, the model generates a personalised time-to-event curve which could aid patient empowerment, triage and service provision. Finally, the problem of defining progression after the CXL operation was investigated. Rather than using fixed thresholds which do not account for the heteroscedasticity within corneal measurements, a variation of Bland Altman analysis was used to generate adaptive thresholds for Kmax, Front K2 and Back K2 which, when combined, constitute a new definition of post CXL progression. This novel method was compared and contrasted with existing definitions of progression using the Kaplan-Meier estimator and could provide a more reliable estimate of keratoconus progression after CXL.
0.1 Impact Statement

Keratoconus is a leading cause of visual impairment in adolescents and young adults, second only to myopia and simple astigmatism, at an age when blurred vision has a maximal effect on the quality of life and employment opportunities. The prevalence of keratoconus varies from 1 in 375 in North Europe [1], to as high as 1 in 48 in some ethnic groups [2, 3].

The research within this thesis has been conducted in close collaboration with several clinicians whose area of expertise is treating patients with keratoconus. The topics investigated were thus carefully selected based on their potential clinical applications. It covers two highly relevant areas within keratoconus care: early detection and progression.

Early detection of keratoconus is a public health priority, and the development of new screening techniques has the potential to make the diagnostic process more efficient. Chapter 2 represents a comprehensive review that will enable researchers, clinicians, and public health policymakers to understand the current state of subclinical keratoconus detection algorithms and provides guidance for future research. Chapter 3 lays important groundwork for future detection algorithms by investigating the use of raw images. If applied to the problem of detecting subclinical disease, this technology could be integrated into screening programs capable of recognising the nascent stages of the disease and thus preventing sight loss.

Keratoconus is a progressive disease and as such, investigating detection alone is not sufficient; it is equally important to be able to understand the prognosis. Chapter 4 presents a prognostic model capable of plotting the trajectory of keratoconus progression based on patients’ data at their first appointment. Personalised modelling of risk may improve patients’ understanding of their condition and the need for CXL. In addition, such a model may aid clinical decision-making to achieve better outcomes and more efficient use of healthcare resources.

Once the CXL operation has been performed, patients require continued monitoring to assess the success of the intervention. It is known that the repeatability of keratometry is reduced at more advanced stages of disease [4, 5, 6, 7], but existing
definitions rely on fixed thresholds. Chapter 5 defines personalised thresholds for post-operative progression which are dependent on an individual’s corneal measurements. Adopting personalised progression thresholds would facilitate more accurate identification of post-operative keratoconus progression, particularly in early disease when there are smaller changes in keratometry.

The content of this thesis has been published in two peer-reviewed Pubmed-indexed journals (cited six times as of January 2023), presented at multiple conferences (IoO ECR Symposium 2021, CXL Experts 2021, Asia Pacific Tele-Ophthalmology Society, European Society of Cataract and Refractive Surgery 2022, Bowman Club 2022) as well as a recorded seminar given to the Keratoconus Group charity (https://www.keratoconus-group.org.uk/index.php/2021/09/25/talk-by-howard-maile/). Furthermore, a web tool (http://beta.moorfieldscxl.com) has been developed and is undergoing evaluation at Moorfields Eye Hospital by clinical staff.
0.2 Declaration

The work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.
0.3 Acknowledgements

I would like to thank the following individuals and credit their specific contributions:

• My primary supervisor, Stephen Tuft, for his research guidance and contributions to chapters 2, 4 and 5.

• My secondary supervisor, Nikolas Pontikos, for his assistance with all chapters of this project and encouraging me throughout.

• Daniel Gore, Olivia Li and Marcello Leucci for their invaluable collaboration and clinical insight, particularly with chapters 2, 4 and 5.

• Alison Hardcastle and Bruce Allan for their suggestions and input into several chapters.

• Lynn Kandakji for her help with annotations, software development and proofreading, particularly in chapter 3.

• Patrick Royston, Mary Fortune and Catey Bunce for their expert statistical knowledge.

• Bart Liefers for his help with segmentation models in chapter 3.

• Ismail Moghul for his advice on software development.

• Siegfried Wagner for his advice on survival analysis.

• Alan Da Silva Sousa Wilter for working on the software behind the annotations process in chapter 3.
0.4 Publications

Several chapters from this thesis have been published. The systematic review from chapter 2 has been published in JMIR Medical Informatics [8]. Chapter 4 has been published in Ophthalmology [9]. At the time of writing, chapter 3 has been submitted as an abstract for the Association for Research in Vision and Ophthalmology (ARVO) 2023 conference and chapter 5 has been submitted for publication.
Contents

0.1 Impact Statement .............................................. ii
0.2 Declaration ..................................................... iv
0.3 Acknowledgements ........................................... v
0.4 Publications .................................................... vi

1 Introduction ...................................................... 1
  1.1 Background .................................................. 1
  1.2 Diagnostic Tools ............................................ 2
    1.2.1 Placido Disc Topography ................................. 2
    1.2.2 Slit-Scanning ......................................... 3
    1.2.3 AS-OCT .............................................. 3
    1.2.4 Scheimpflug Imaging .................................. 5
    1.2.5 Applanation Tonometry ................................. 7
  1.3 Diagnostic Data Types ....................................... 7
    1.3.1 Refraction ............................................. 8
      1.3.1.1 Subjective Refraction .............................. 8
      1.3.1.2 Objective Refraction ............................... 8
    1.3.2 Corneal Imaging Data and Derived Parameters ........ 9
    1.3.3 Biomechanical Data .................................. 11
    1.3.4 Other Non-Cornea Specific Parameters ............... 11
  1.4 Grading Systems and Indices ................................ 12
  1.5 Treating Keratoconus with Collagen Cross-linking ........ 12
  1.6 Computational Methods .................................... 13
    1.6.1 Machine Learning .................................... 13
2 Subclinical Keratoconus Detection Literature Review

2.1 Introduction

2.1.1 Case definition for keratoconus

2.1.2 Grading Systems for Subclinical Keratoconus

2.2 Methods

2.2.1 Search Strategy

2.2.2 Inclusion and Exclusion Criteria

2.2.3 Data Synthesis

2.3 Results

2.3.1 Selection Process and Results Overview

2.3.2 Subclinical Keratoconus Detection Data Types

2.3.2.1 Aberrometry

2.3.2.2 Corneal Imaging Data and Derived Parameters

2.3.2.3 Biomechanical data

2.3.2.4 Demographic Risk Factors

2.3.3 Machine Learning Techniques

2.3.3.1 Neural Networks

2.3.3.2 Convolutional Neural Networks

2.3.3.3 Decision Trees

2.3.3.4 Random Forests

2.3.3.5 Discriminant Analysis

2.3.3.6 Support Vector Machines

2.3.3.7 Logistic Regression

2.3.3.8 Comparative Studies
2.3.3.9 Unsupervised Learning .......................... 36
2.3.4 Validation ........................................... 37
2.4 Discussion .............................................. 38
  2.4.1 Data Types for Detection of Subclinical Keratoconus ...... 38
  2.4.2 Machine Learning ................................. 39
  2.4.3 Validation ......................................... 39
  2.4.4 Strengths and Limitations ........................... 40
  2.4.5 Challenges and Future Directions .................... 40
  2.4.6 Conclusion ....................................... 43

3 Early versus Late Stage Keratoconus Classification Using Deep Learning 44
  3.1 Introduction ........................................... 45
  3.2 Data .................................................. 45
    3.2.1 Pre-processing ................................. 46
  3.3 Methods ............................................. 47
    3.3.1 Classification Model ......................... 47
    3.3.2 Segmentation Model ............................ 48
    3.3.3 Approach 1: Classification Only .............. 48
      3.3.3.1 Training and Validation .................. 50
    3.3.4 Approach 2: Segmentation Followed by Classification .... 50
      3.3.4.1 Training and Validation .................. 50
    3.3.5 Software ....................................... 52
  3.4 Results .............................................. 52
    3.4.1 Approach 1: Classification Only ............... 52
    3.4.2 Approach 2: Segmentation Followed by Classification .... 55
      3.4.2.1 Segmentation ............................... 55
      3.4.2.2 Classification .............................. 56
  3.5 Discussion .......................................... 59

4 Pre CXL Progression ..................................... 62
  4.1 Introduction ....................................... 63
4.2 Methods ......................................................... 63
  4.2.1 Cohort .................................................. 63
  4.2.2 Model Fitting and Covariate Selection .............. 65
  4.2.3 Keratometric Progression Sensitivity Analysis .... 66
  4.2.4 Multivariable Model Validation ....................... 66
  4.2.5 Statistical Analysis .................................... 67

4.3 Results .......................................................... 69
  4.3.1 Cohort .................................................. 69
  4.3.2 Model Fitting and Covariate Selection (Genetic Data) .... 70
  4.3.3 Model Fitting and Covariate Selection (Excluding Genetic Data) 72
  4.3.4 Keratometric Progression Sensitivity Analysis .... 79
  4.3.5 Multivariable Model Validation ....................... 80
  4.3.6 Multiple Imputation Results ......................... 82

4.4 Discussion ....................................................... 83

5 Post CXL Progression ............................................. 86
  5.1 Introduction ............................................... 87
  5.2 Methods ................................................ 88
  5.3 Data ....................................................... 88
    5.3.1 Adaptive Thresholds ................................. 89
    5.3.2 Post CXL Progression Definition .................. 90
    5.3.3 Statistical Methods ................................ 91
  5.4 Results .................................................... 93
    5.4.1 Adaptive Thresholds ................................. 93
    5.4.2 Post CXL Progression Definition .................. 95
  5.5 Discussion ................................................ 97

6 Discussion ......................................................... 101
  6.1 Detection .................................................. 101
  6.2 Progression .............................................. 103
  6.3 Future Directions ......................................... 104
Acronyms

AI  Artificial Intelligence
AIC  Akaike Information Criterion
AK  Amsler-Krumeich
AS-OCT  Anterior Segment Optical Coherence Tomography
AUC  Area under the ROC Curve
Back K1  Flat posterior keratometry in the central 3 mm zone
Back K2  Steep posterior keratometry in the central 3 mm zone
BAD  Belin / Ambrosio Enhanced Ectasia Display
BFS  Best Fit Sphere
BIC  Bayes Information Criterion
CCT  Central Corneal Thickness
CDS  Clinical Decision Support
CH  Corneal Hysteresis
CNN  Convolutional Neural Network
CRF  Corneal Resistance Factor
CSL  Corneal cross-linking
CXL  Corneal Collagen Cross-Linking

DES  Dry Eye Syndrome

DT   Decision Trees

EKC  Early Keratoconus Clinic

EPR  Electronic Patient Record

ESI  Ectasia Screening Index

ETDRS Early Treatment Diabetic Retinopathy Study

FDA  Food and Drug Administration

FFKC Forme Fruste Keratoconus

Front K1 Flat anterior keratometry in the central 3 mm zone

Front K2 Steep anterior keratometry in the central 3 mm zone

GWAS Genome-wide Association Study

HOA  Higher-Order Aberration

HR   Hazard Ratio

IECV Internal-external Cross Validation

IHA  Index of Height Decentration

ISV  Index of Surface Variance

IVA  Index of Surface Asymmetry

K1   Flat Keratometry

K2   Steep Keratometry

LoA  Limits of Agreement
Kmax  Maximum Keratometry
Kmin  Minimum Keratometry
KMI   Keratoconus Match Index
LASIK Laser Assisted In Situ Keratomileusis
LoA   Limits of Agreement
LOA   Lower-Order Aberration
LR    Logistic Regression
MAE   Mean Anterior Elevation
MFP   Multivariate Fractional Polynomial
MEH   Moorfields Eye Hospital
ML    Machine Learning
MPE   Mean Posterior Elevation
NN    Neural Network
OCT   Optical Coherence Tomography
ORA   Ocular Response Analyzer
RF    Random Forest
RNN   Recurrent Neural Network
SD-OCT Spectral Domain Optical Coherence Tomography
SNP   Single-nucleotide Polymorphism
SVM   Support Vector Machines
TBI   Tomography and Biomechanical Index
**TD-OCT**  Time Domain Optical Coherence Tomography

**TKC**  Topographic Keratoconus Classification

**TP**  Thinnest Point

**VAE-E**  Very Asymmetric Ectasia-Ectatic

**VAE-N**  Very Asymmetric Ectasia-Normal

**VAE-NT**  Very Asymmetric Ectasia-Normal Topography
Chapter 1

Introduction

1.1 Background

Keratoconus is a bilateral ectatic disease of the cornea that can cause visual loss through corneal distortion and scarring [10, 11]. Figure 1.1 depicts both a normal cornea and a keratoconic cornea. Keratoconus is a leading cause of visual impairment in adolescents and young adults, second only to myopia and simple astigmatism, at an age when blurred vision has a maximal effect on quality of life and employment opportunities. The prevalence of keratoconus varies from 1 in 375 in North Europe [1], to as high as 1 in 48 in some ethnic groups [2, 3]. The onset of disease is typically after puberty, with subsequent progression at a variable rate over two to three decades. As the disease advances, corneal distortion can reach a stage where spectacle-corrected vision is inadequate and patients must rely on soft or rigid contact lenses to achieve good functional vision. However, contact lenses are not always tolerated and the visual impairment can severely affect quality of life [12, 13]. In the natural course of the disease, around 20% of patients will be offered a corneal transplant to improve their vision but at the risk of postoperative complications (e.g. microbial keratitis and inflammation), potential allograft rejection and transplant failure [14, 15]. The majority of individuals with keratoconus are identified because of symptoms of visual disturbance or an increase in astigmatism at refraction. It is therefore inevitable that most individuals with keratoconus are detected at a stage
when visual deterioration has already occurred.

Figure 1.1: Diagram showing both a normal cornea (left) and an ectatic cornea suffering from keratoconus (right). The keratoconic cornea can be seen to be bulging and thinning compared to the normal cornea. Diagram courtesy of [16].

1.2 Diagnostic Tools

There is a wide range of hardware and software available for diagnosing keratoconus. In the section below, only the most important tools or techniques relevant to this thesis are described.

1.2.1 Placido Disc Topography

Placido disc topography is used to measure the shape of the cornea. It projects a series of concentric rings onto the cornea and captures the reflection using a camera. By examining the reflections of the rings, a topographical map of the anterior surface of the cornea can be generated. Placido disk topography cannot analyse the posterior surface of the cornea, which is useful when diagnosing keratoconus [17]. In order to do this, tomographic imaging techniques such as AS-OCT or Scheimpflug imaging must be employed. Tomographic devices such as the MS-39 (CSO, Firenze, Italy) or Sirius (CSO, Firenze, Italy) combine tomography with Placido disc topography.
1.2.2 Slit-Scanning

Slit-scanning is an imaging technique which involves sequentially projecting 40 slits of light (20 from the left and 20 from the right) at a fixed angle to the cornea in order to create an overlapping pattern. By analysing the reflection pattern, a topographical map can be created. The Orbscan II (Bausch and Lomb Inc., Rochester, New York, USA) is an example of a slit-scanning device [18].

1.2.3 AS-OCT

Optical Coherence Tomography (OCT) was first developed for imaging the posterior segment of the eye and was subsequently extended to imaging the anterior segment where it is known as AS-OCT. The basic idea behind OCT is analogous to ultrasound; a beam of light is directed at the anterior surface of the eye and the timing and intensity of the reflected waves is measured from which the depth and tissue type can be inferred. In order to take these measurements in practice, an identical reference beam is reflected from a scanning mirror with known path length.

![AS-OCT Diagram](image)

**Figure 1.2:** AS-OCT Diagram courtesy of [19].

A fundamental principle of OCT is low-coherence interferometry. Interference between two beams of low-coherence light only occurs when the difference in optical path lengths is within a very small range. This allows for very high-resolution measurements in the order of micrometres. In order to take a single depth scan, the reference mirror is scanned through a range of depths whilst the sample remains fixed.
The light intensity resulting from the recombination of the two beams is measured throughout the depth scan from which a depth profile can be created. The process is repeated by moving to the next lateral position leading to a cross-sectional image of the anterior segment. This type of OCT is known as Time-Domain OCT (TD-OCT). Another form of OCT is Spectral Domain OCT (SD-OCT) or Spectrometer-based OCT. The basic principle is the same but rather than moving the reference mirror to generate a depth scan, the reference mirror remains in a fixed position. The spectrum of light generated by the recombination of the light beams is measured by a spectrometer which is then Fourier transformed in order to generate a depth profile [20]. Figure 1.2 depicts both TD-OCT and SD-OCT.

There are multiple manufacturers of AS-OCT devices used in the diagnosis of keratoconus. In chapter 3, the MS-39 device is used. The MS-39 device combines Placido disk corneal topography with SD-OCT. It is able to provide information on pachymetry, elevation and keratometry and dioptric power either as heat maps or raw numeric data. In addition, the raw images can be downloaded and examined. Example heat maps and a raw image from the MS39 device can be seen in figures 1.3 and 1.4 respectively.

**Figure 1.3:** Example heat maps for various corneal parameters captured from the MS-39 device.
1.2.4 Scheimpflug Imaging

When acquiring a corneal image, there are three planes of interest: the object plane, the lens plane and the image plane. When these three planes are parallel, an image can be captured in its entirety with perfect focus. When the object plane becomes tilted (as is expected when imaging the cornea due to its curving nature), the image will not be entirely focused. To resolve this, the Scheimpflug principle is applied; the image plane is rotated to a specific angle (dictated by the Scheimpflug intersection) whereby it is possible to bring the entire image into focus [17]. Figure 1.5 depicts the Scheimpflug principle. Common examples of Scheimpflug devices include the Pentacam (Oculus GmbH, Wetzlar Germany) and Sirius. In chapters 4 and 5, the data is acquired from the Pentacam. Figure 1.6 shows an example Pentacam heat map for an eye with advanced keratoconus.
**Figure 1.5:** Scheimpflug principle; (a) All planes are parallel; the image is focused (b) Object and image plane are not parallel; the image is not entirely focused. (c) The object and image plane are at different orientations but the image plane has been rotated according to the Scheimpflug principle, resulting in a fully focused image. Diagram and text courtesy of [17].

**Figure 1.6:** Heat maps of an advanced keratoconus eye derived from Scheimpflug corneal imaging using the Pentacam device.
1.2.5 Applanation Tonometry

![Diagram illustrating the acquisition process using the ORA. A collimated beam of Infrared (IR) light is directed at the cornea and a light detector measures the collimation of the reflected beam. A jet of air is directed at the cornea which causes its surface to deform from convex to concave and back to convex. The pressure at which the reflected light becomes collimated is recorded. This occurs twice, once when the cornea moves from a convex shape to an applanated state (flat), and once again when it moves from a concave shape to an applanated state. Image courtesy of [21].](image)

**Figure 1.7:** Diagram illustrating the acquisition process using the ORA. A collimated beam of Infrared (IR) light is directed at the cornea and a light detector measures the collimation of the reflected beam. A jet of air is directed at the cornea which causes its surface to deform from convex to concave and back to convex. The pressure at which the reflected light becomes collimated is recorded. This occurs twice, once when the cornea moves from a convex shape to an applanated state (flat), and once again when it moves from a concave shape to an applanated state. Image courtesy of [21].

Applanation Tonometry measures the intraocular pressure (IOP) and other biomechanical properties thought to be useful in diagnosing keratoconus such as corneal hysteresis (CH) and corneal resistance factor (CRF) by measuring the force required to flatten the cornea. The most advanced device for this is the Ocular Response analyser (ORA), Reichert Ophthalmic Instruments, Buffalo, NY). It uses a jet of air directed at the cornea and measures the deformation response. Figure 1.7 illustrates this process. Another device capable of applanation tonometry is the Corvis ST (Oculus GmbH, Wetzlar Germany), which uses a high-speed Scheimpflug camera to measure distortion on cross-sectional images.

1.3 Diagnostic Data Types

In this section, the four main groups of data used to diagnose keratoconus are explained. Figure 1.8 summarises these in an organisational diagram.
1.3.1 Refraction

1.3.1.1 Subjective Refraction

Subjective refraction refers to an eye examination where there is cooperation between the patient and clinician. The most common example of this is the Snellen visual acuity test where a patient is seated six metres from the chart and asked to read the smallest letters possible. As keratoconus progresses, visual acuity will decrease. As such, visual acuity is useful both in diagnosing keratoconus and also severity grading [22].

1.3.1.2 Objective Refraction

Objective refraction obtains measurements of refractive error without input from the patient. When diagnosing keratoconus, the most useful form of objective refraction is aberrometry. Aberrations are produced by imperfections in the optical quality of the refracting surface of the eye including the cornea and the lens. They can be described as a set of Zernike polynomials or with Fourier analysis. With the Zernike method, aberrations can be sub-classified as lower-order aberrations (LOA) and higher-order aberrations (HOA). LOAs include simple defocus (myopia or hyperopia) and regular astigmatism, which accounts for approximately 90% of the refractive error of the normal eye [23]. The most clinically relevant HOAs are spher-
ical aberration, coma and trefoil that cannot be corrected by glasses or a soft contact lens. In keratoconus the irregular distortion of the front and back surfaces of the cornea causes visually significant HOAs. HOAs are measured from the distortion of a plane wavefront of light passing through the optics of the eye. However, HOAs can also be derived indirectly from measurement of any distortion (e.g. elevation) of the corneal surfaces.

1.3.2 Corneal Imaging Data and Derived Parameters

As explained in section 1.1, there are various acquisition techniques including Scheimpflug optics, AS-OCT, and horizontal slit-scanning systems. These systems incorporate software that processes the images to derive numerical indices or secondary images that represent various aspects of corneal shape. The main categories of output from these devices are measurements of the corneal surface radius of curvature (keratometry), elevation or depression of a point on the corneal surface from the mean (elevation), corneal thickness (pachymetry) and displacement. A schematic diagram explaining the concepts behind these parameters can be seen in figure 1.9.

**Keratometry parameters** Keratometric parameters measure the radius of curvature of the anterior or posterior corneal surfaces. Examples include the meridian with the steepest corneal radius of curvature (corresponding to $K_{\text{max}}$) and the flattest curvature (corresponding to $K_{\text{min}}$). See diagram E in figure 1.9.

**Elevation parameters** Elevation represents points above or below the best fit sphere (BFS) of the corneal surface measured in microns. For the posterior cornea, this is measured either as the divergence from the best-fit of the whole posterior corneal diameter, or divergence from the best-fit of an annulus of the peripheral posterior corneal surface outside the central 4mm [24]. The latter method refers to the Belin-Ambrosio map, and better describes the central corneal elevation known to keratoconus. Values can be presented as either colour-coded maps, individual pa-
Figure 1.9: Schematic diagram illustrating the four basic types of corneal parameters. (A) pachymetry: distance between anterior and posterior corneal surfaces, (B) displacement: distance between the apex of the cornea and the point of minimum thickness. (C) and (D) depict two methods of calculating elevation using the Best Fit Sphere (BFS). In (C) the BFS is fitted to both the normal peripheral posterior surface (blue) and the abnormal anterior protrusion of the central posterior surface (green). In (D) the BFS is fitted to only the normal peripheral posterior surface (blue) excluding the abnormal central posterior surface (green). (E) keratometry: the smallest radius corresponds to the largest refractive power (Kmax) and the largest radius corresponds to the smallest refractive power (Kmin).

Displacement parameters These represent measures such as the displacement of the point of minimum corneal thickness from the corneal apex. See diagram B in figure 1.9.

Pachymetry parameters Pachymetry is the thickness of the cornea measured either with ultrasound or imaging techniques. Simple examples of these include cen-
tral corneal thickness (CCT) and the thinnest point of the cornea. A reduction in corneal thickness is a fundamental biomarker of keratoconus. See diagram A in figure 1.9.

**Summary indices** In addition to single-parameter measurements (CCT, etc), tomographic systems such as the Pentacam can also combine measurements to compute derived indices that estimate the regularity of corneal shape. Basic indices such as the index of surface variance (ISV), index of vertical asymmetry (IVA) or index of height asymmetry (IHA) are formed from multiple data points. Composite indices are themselves formed from other indices and data points. Examples of these include Keratoconus Index (KI), KISA % and BAD-D. Further information regarding indices can be found in section 1.4.

**Heat Maps** Modalities such as Scheimpflug and AS-OCT capture images at various corneal meridians and subsequently use this data to derive the heat maps that facilitate visual interpretation of the data, although there is extrapolation of the data in areas between the imaged meridians. For example, the Pentacam can translate the raw images into several types of colour heat maps (e.g. axial curvature, posterior/anterior elevation, regional pachymetry), all based on the same original tomography dataset.

**1.3.3 Biomechanical Data**

Corneal biomechanics refers to the distortion response of the cornea to an applied force. As mentioned in section 1.2.5, the ORA is an example of a device capable of capturing biomechanical data such as CH and CRF. However, there is disagreement as to its utility in the diagnosis of keratoconus [25, 26].

**1.3.4 Other Non-Cornea Specific Parameters**

Other information useful when detecting keratoconus includes genetic, demographic and associated disease data. A recent study large scale GWAS identified 36 genetic
associated with the development of keratoconus [27]. Studies have shown that certain demographic factors, such as age and ethnicity, may influence the prevalence and severity of keratoconus [28, 29]. Additionally, Down syndrome and ocular allergies are known risk factors in the development of keratoconus [29].

1.4 Grading Systems and Indices

There are a number of clinical systems to grade the severity of keratoconus, but an expert consensus report in 2015 concluded that none were clinically validated [29]. A frequently used system is Amsler-Krumeich (AK) grading, which is based on keratometry, pachymetry, astigmatism and the presence or absence of scarring [30] and classifies the severity of keratoconus into four stages. In chapter 3, stage 1 (early) versus stage 4 (late) keratoconus is classified using a deep learning algorithm and the ground truth classification uses similar keratometry thresholds to the AK grading system. Another common grading system is the ABCD grading system which is based on keratometry, pachymetry and visual acuity which again classifies into four stages [31].

Other grading systems based on the Pentacam include the topographic keratoconus classification (TKC) [32]. When considering other imaging systems, the ORA’s keratoconus match index (KMI) is based on a composite value of seven parameters derived from the reflected waveform generated as the cornea is distorted by a jet of air [33].

1.5 Treating Keratoconus with Collagen Cross-linking

Collagen cross-linking is a procedure used to treat corneal ectasia such as keratoconus. It is a photochemical treatment of the cornea with UV-A light following application of riboflavin (vitamin B2), which can arrest progression of keratoconus in 98.3% of eyes, even in relatively advanced disease [35, 36, 37, 38, 39, 40]. Figure 1.10 depicts this process. The procedure causes the collagen fibres in the cornea to
bond (or cross-link), which results in stiffening and this helps to slow or stop the progressive thinning and bulging exhibited by keratoconic corneas. It is normally performed as an out-patient procedure and takes between 30 and 60 minutes.

1.6 Computational Methods

The following sections introduce and explain the prerequisite computational methods used within this thesis.

1.6.1 Machine Learning

Chapters 2 and 3 make reference to a variety of machine learning algorithms in the context of keratoconus detection. Machine learning is a branch of artificial intelligence centred on writing software capable of learning from data in an autonomous fashion by minimising a loss function or maximising the likelihood [41]. It can be broadly classified as either supervised or unsupervised learning [42]. In supervised learning, the algorithm is trained with input data labelled with a desired output so that it can predict an output from unlabelled input data [43]. In comparison, in unsupervised learning, the algorithm is not trained using labelled data. Instead, the algorithm is used to identify patterns or clusters in the data [44]. When applied to the field of keratoconus detection, machine learning may be used to analyse a
large number of corneal parameters that can be derived from corneal imaging as well as other clinical and biometric measures such as visual acuity and refraction to predict the disease [45]. It can also be applied directly to imaging data to work at the pixel level [46]. Deep learning, a specific branch of machine learning, uses artificial neural networks (NNs) with multiple layers to process input data [47]. It is particularly well suited to the segmentation or classification of corneal images [48]. Both machine learning and deep learning may facilitate superior diagnostic ability that, when implemented as automated screening tools, could result in significant advances in case detection, mitigating both the cost of new imaging hardware and the burden on ophthalmic health care professionals [49]. In addition, through unsupervised learning, it may be possible to discover previously unknown disease subtypes or features [50, 51].

### 1.6.2 Image Classification

The problem of image classification lends itself well to CNNs. At the heart of a trained CNN are various layers of image filters known as convolutional layers. In the field of traditional computer vision, image filters are small matrices that are applied to pixels within an image using a mathematical operation known as convolution. When image filters are convolved with every pixel in the image, they can be used for tasks like sharpening, blurring and edge detection and the desired result is controlled by the values within the image filter. In the context of CNNs, the exact values within the image filters are known as weights and are iteratively learned using the same fundamental techniques as regular neural networks. CNNs usually consist of multiple layers of filters known as convolutional layers which are used successively to extract hierarchical features from the image. Between each convolutional layer is an activation function and a pooling layer. The activation function is responsible for deciding which information is passed between layers and aids the training process by introducing non-linearity. The pooling layer reduces the dimensions of the data and summarises the features learned from the previous layer. The final part of the CNN flattens the output from the previous layers into a one-dimensional vector.
and uses a separate activation function to classify the output. At this point, a loss function is used to calculate the difference between the ground truth and the prediction. This loss is backpropagated through the network to adjust the filters (weights) within each convolutional layer and the process is repeated until a satisfactory loss is achieved [44]. A full pass through the image set is known as an epoch and usually the CNN is trained for a certain number of epochs or until no further improvement in the loss is observed.

1.6.3 Image Segmentation

![Raw image](image1.png) ![Mask](image2.png)

**Figure 1.11:** Example input image (a) and mask (b) for training a segmentation model

Image segmentation refers to the task of separating an image into subgroups called image segments. The most popular algorithm for segmentation is U-Net, which is a special class of convolutional neural network. U-Nets have two sections. The first section is similar to a normal CNN in which downsampling is achieved by pooling operators. The second path does the opposite: upsampling layers increase the resolution of the output to the original image size. Skip connections relate the information across the downsampling and upsampling process. Intuitively, the first section
is responsible for classification and the second section is responsible for localisation of this information. A diagram can be seen in figure 1.12. The input is the image we wish to segment and the output is a segmentation map. In order to train a U-Net, we need both the input image we wish to segment and also an image mask which serves as the label. Examples of these for a cornea can be seen in figures 1.11(a) and 1.11(b).

When evaluating the accuracy of a segmentation model, the most commonly used metric is the Dice Similarity Coefficient (dice score) which compares pixels between both the ground truth mask and the predicted mask. The formula for this is shown below:

\[
\text{Dice} = \frac{2TP}{(2TP + FP + FN)}
\]

Where \(TP, FP, FN\) refer to the number of true positives, false positives and false negatives respectively.

Figure 1.12: Typical U-Net architecture courtesy of Ronneberger et al. 2015 [52]
1.6.4 Survival Analysis

Survival analysis is a branch of statistics focusing on the analysis of time-to-event data. It commonly involves estimating the survival function along with hazard ratios.

The survival function describes the proportion of the population who have not yet experienced the event by time t. The most simple method to estimate it is with the Kaplan-Meier estimator:

\[
S(t) = \prod_{i: t_i \leq t} (1 - \frac{d_i}{n_i})
\]

where \(S(t)\) is the survival function, \(t_i\) is the time when at least one event happened, \(d_i\) is the number of events at time \(t_i\) and \(n_i\) is the number of individuals that have survived up to time \(t_i\).

A major limitation of the Kaplan-Meier estimator, however, is that it cannot be used for multivariate analysis and cannot handle continuous covariates. When this form of analysis is required, researchers generally turn to the Cox Proportional Hazards model which is essentially a regression model used to investigate the effect of multiple covariates on the survival time. The general equation of the Cox Proportional Hazards model is as follows:

\[
h(t) = h_0(t) \exp \beta X
\]

where \(h(t)\) is the hazard function, \(h_0(t)\) is the baseline hazard function, \(\beta\) is a vector of covariate coefficients and \(X\) is a vector of covariates. It is worth noting that the survival function can be derived from the hazard function.

The cox proportional hazard model is a commonly used method of generating hazard ratios. However it does not allow the estimation of smooth time-to-event curves. In the analysis in chapter 4, a Royston-Parmar model is used, which performs direct estimation of the hazard function [53] and thus facilitates the generation of smooth curves. The equation for the Royston-Parmar model is commonly written in terms of the log cumulative hazard function:

\[
\ln H(t)|x = s(\ln t\gamma) + \beta X
\]
where $H(t)$ is the cumulative hazard function, $s(\ln t \gamma)$ is a restricted cubic spline that is a function of derived variables $\gamma$, $\beta$ is a vector of covariate coefficients and $X$ is a vector of covariates.

1.6.5 Discrimination and Explained Variation

The primary measures used in chapter 4 for assessing model performance are discrimination and explained variation. Discrimination refers to the ability of the model to distinguish between patient survival probabilities. For example, a survival model that has a wide range of predicted survival probabilities at 2 years is superior to a model with a narrow range [53]. Explained variation ($R^2$) refers to the proportion of the total variance in the data that is explained by the model or explanatory variable. It can also be viewed as a measure of goodness of fit. Chapter 4 uses variations of discrimination and explained variation appropriate for evaluating Royston Parmar models known as Royston and Sauerbrei’s D statistic and Royston and Sauerbrei’s $R^2_D$ respectively [54].

1.7 Detection

Diagnosing advanced or clinical grade keratoconus can be achieved using the diagnostic tools listed above either singly or in combination. According to the Global Consensus on Keratoconus and Ectatic Diseases [29], the diagnosis of keratoconus requires the following mandatory findings:

- Abnormal posterior elevation
- Abnormal corneal thickness distribution
- Clinical noninflammatory corneal thinning

The most important stage of keratoconus for identification is very early stage keratoconus (subclinical keratoconus). By detecting subclinical keratoconus and providing early treatment, clinicians may be able to improve visual outcomes and quality of
life for patients. The detection of keratoconus at an earlier stage has become increasingly relevant since the introduction of corneal collagen cross-linking (CXL) [55]. Paediatric keratoconus cases often progress faster than adults which further highlights the need for early detection [56]. The benefit of early intervention to minimise visual loss is clear, and there is evidence that it is cost-effective [57, 58, 59], but the mechanism for improving early diagnosis by community-based optometrists is challenging because asymptomatic patients with subclinical disease are unlikely to seek review [60]. Keratoconus is still diagnosed at a relatively late stage and despite advances in imaging equipment, more research into screening methods is required to develop a cost effective solution [55].

In light of the importance of early detection, chapter 2 of this thesis provides a comprehensive literature review of existing algorithms for detecting subclinical keratoconus. Chapter 3 presents the results from a new algorithm using deep learning applied to raw AS-OCT images to classify early versus late stage keratoconus, which should be viewed as the precursor to future work.

1.8 Progression

Keratoconus is a progressive disease and, as such, detecting signs of progression is important for early diagnosis and it is well known that younger patients and paediatric cases tend to be faster progressors [61, 56]. According to the Global Consensus of Keratoconus and Ectatic Diseases, progression is defined by a consistent change in at least two of the following parameters: steepening of anterior corneal surface, steepening of posterior corneal surface, thinning and/or an increase in the rate of corneal thickness change [29]. Increasing Kmax is the most commonly used parameter for defining steepening of the corneal surface [62]. However some authors believe that Kmax is an unreliable parameter for detecting progression because keratometry has been shown to have poor repeatability in established keratoconus patients [63, 64, 65, 66]. Other measures include visual acuity and pachymetry, but they have been found to be poor indicators of keratoconus progression [67]. In ad-
dition, monitoring the anterior cornea surface with topometric indices such as ISV and IHD has been put forward as the best performing indicators of progression [68].

Depending on the definition chosen and the rate of progression, these thresholds may be passed in a few months, years, or not at all. At the first assessment it can be a challenge to distinguish eyes that are at risk of rapid progression from those where it is safe to monitor. Furthermore, unnecessary review visits are a burden to the patient and the care system. Chapter 4 attempts to address this by presenting a statistical model for predicting the probability of receiving CXL given corneal parameters available at the first appointment. The problem of recognising progression is not limited to before the CXL operation. Chapter 5 explores new methods for defining progression after CXL and compares and contrasts these with existing methods. The EKC (Early Keratoconus Clinic) has a large amount of readily available historical appointment data, thus the related demographic and serial tomography data from this cohort of patients was used for analysis.
Chapter 2

Subclinical Keratoconus Detection

Literature Review
2.1 Introduction

This chapter presents a systematic review of the literature surrounding early detection of keratoconus using computational methods, in particular machine learning. Advanced keratoconus is relatively easily diagnosed clinically and therefore developing machine learning algorithms to identify advanced disease has limited utility. The literature review was thus directed toward publications that included detection of subclinical keratoconus because identifying these individuals would allow for early intervention with CXL to reduce the likelihood of disease progression. The review is structured both around the different types of available input data (parameters, indices and corneal imaging systems) and the machine learning algorithms for keratoconus detection.

2.1.1 Case definition for keratoconus

A number of terms describe the early stage of keratoconus prior to vision being affected, including “forme fruste” keratoconus (FFKC), keratoconus suspect, subclinical keratoconus, and preclinical keratoconus. The most commonly used terms are FFKC and subclinical keratoconus but there is no consensus on their definition [69]. Because of the overlap in the nomenclature, with little evidence of which, if any, poses a particular risk for progression to clinical keratoconus, all papers that contain an identifiable subgroup of eyes with any of these definitions were included for review. Papers that only consider eyes with established keratoconus were excluded.

2.1.2 Grading Systems for Subclinical Keratoconus

Unlike diabetic retinopathy, where there is a widely adopted diagnostic grading system (ETDRS) [70] and where diagnosis of early disease is based on the presence of discrete entities on the retina (e.g. microaneurysms), the diagnostic grading of subclinical keratoconus has not yet reached the same level of consensus [29]. Frequently used grading systems such as AK and ABCD do not specifically include a grade for subclinical keratoconus although the ABCD grading system has a grade
0 for ‘probable absence of disease’. Placido disc-based topographic indices include Pathfinder Corneal Analysis, keratoconus percentage index (KISA%), KPI and KCI indices, and I-S value [71]. These indices are based on anterior elevation and pachymetry and thus may miss earlier signs of keratoconus such as posterior elevation change and horizontal displacement of the zone of thinning [71].

The tomography based Belin/Ambrósio Enhanced Ectasia Display (BAD) is an inbuilt function in the Pentacam designed to generate a comprehensive display for keratoconus and ectasia screening through the analysis of multiple variables [71]. Regression analysis using a database of parameters from normal individuals and patients with keratoconus generates a numeric ‘D’ value (BAD-D), which indicates how the examined cornea varies from ‘normal’. However, it was not specifically designed to identify subclinical keratoconus, and has 81.1% sensitivity and 73.2% specificity for distinguishing suspect keratoconus from controls [72].

2.2 Methods

2.2.1 Search Strategy

The literature review was conducted upon the evidence for the utility of machine learning applied to the detection of keratoconus published between Jan 1, 2010 and Oct 31, 2020. The PRISMA Statement 2009 criteria [73] was followed to search four bibliographic databases: MEDLINE, EMBASE, Web of Science and Cochrane using keyword search on their title, abstract and keywords (textbox below). The review was not registered and no protocol was prepared.

```latex
((keratoconus) OR (cornea* protru*) OR (cornea* ectasia)) AND ((algorithm) OR (machine learn*) OR (deep learn*) OR (artificial intelligence) OR (detect*) OR (diagnos*) OR (screen*) OR (examin*) OR (analys*) OR (investigat*) OR (identif*) OR (discover*) OR (interpret*) OR (test*))
```
2.2.2 Inclusion and Exclusion Criteria

Only studies that investigate the detection of early keratoconus, or investigate a subgroup of patients with early disease were included, as defined by one of the following terms: subclinical keratoconus; forme fruste keratoconus (FFKC); preclinical keratoconus; suspect keratoconus; unilateral keratoconus (normal fellow eye); very asymmetric ectasia (normal fellow eye); and any definition considered to be equivalent to the aforementioned terms. The studies should report the performance of their model on a dataset that was separate from the training dataset (often called a validation or a test set). This includes splitting the dataset into training and test sets (e.g. 70% training, 30% testing); K-fold cross-validation (an extension of simple splitting but the process is repeated K times, e.g. when K=10, partition the dataset into 90% for training the model and 10% for testing, and the process is repeated 10 times by choosing a different 10% partition each time for testing); or evidence of validation study where the aim is to assess a previously derived model on a new dataset (also known as an external validation). Finally the full text article should be available and only papers published in English were considered.

Papers based on the detection of ‘early’ keratoconus defined as Amsler-Krumeich stages 1 or 2 as this represents established KC with both clinical and topographical features [74] were excluded. Additionally, manuscripts which did not perform any type of validation on their results were excluded.

2.2.3 Data Synthesis

Two reviewers (HM and JOL) screened the initial results based on the inclusion criteria. These results were then screened for the exclusion criteria by HM and NP. Any disagreements in meeting inclusion or exclusion criteria were resolved by discussion. Once the set of articles was finalised, two reviewers (HM and JOL) analysed each article and extracted the following information into a master table available in Multimedia Appendix 2 from the published text [8]: author and year, title, system, sample source, country, age, gender, number of eyes for each group, diagnosis details, validation details, input details, input types, method, classification groups,
sensitivity, specificity, accuracy, precision, area under the receiver operating characteristics curve (AUC) and source code availability. The most important information for all results are summarised in table 2.1. The main effect measures sought were sensitivity and specificity. If these statistics were not directly available from the article, they were calculated manually using their standard definition [75].

2.3 Results

2.3.1 Selection Process and Results Overview

1998 potentially relevant papers published between 2010 and 2020 were identified. After filtering, 26 articles were included in the qualitative analysis. Table 2.1 summarises these results and a more extensive version can be found in Multimedia Appendix 2 from the published text [8]. In section 2.3.2, the results are discussed in terms of their input data. The organisation follows the same structure as found in the introduction chapter (figure 1.8). In section 2.3.3, the results were considered in terms of machine learning algorithms and figure 2.1 presents an organisational diagram of the relevant machine learning algorithms. In an effort to maintain consistency, the term ‘subclinical keratoconus’ was used throughout, regardless of the nomenclature used by the original authors. The original term will be included in parenthesis and details of the exact definition can be found in Multimedia Appendix 2 from the published text [8].

2.3.2 Subclinical Keratoconus Detection Data Types

2.3.2.1 Aberrometry

Aberrometry was used to detect subclinical keratoconus in 8 out of 26 papers [76, 78, 80, 100, 83, 89, 93, 99]. Arbelaez et al. (2012) used HOA parameters in their subclinical keratoconus detection model and included a weighted sum of HOAs (known as the Baiocchi-Calossi-Versaci index) and the root mean square of HOAs [76]. Five other studies also used derived Zernike aberrometry data.
<table>
<thead>
<tr>
<th>Author and Year</th>
<th>Number of Eyes</th>
<th>Results (%)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arbelaez et al. 2012[76]</td>
<td>1259</td>
<td>426</td>
<td>92</td>
<td>97.7</td>
</tr>
<tr>
<td>Saad et al. 2012[77]</td>
<td>69</td>
<td>34</td>
<td>92</td>
<td>96</td>
</tr>
<tr>
<td>Smadja et al. 2013[78]</td>
<td>177</td>
<td>47</td>
<td>93.6</td>
<td>97.2</td>
</tr>
<tr>
<td>Ramos-Lopez et al. 2013[79]</td>
<td>50</td>
<td>24</td>
<td>33</td>
<td>78</td>
</tr>
<tr>
<td>Cao et al. 2014[60]</td>
<td>39</td>
<td>49</td>
<td>94</td>
<td>90</td>
</tr>
<tr>
<td>Buhren et al. 2014[80]</td>
<td>245</td>
<td>32</td>
<td>78.1</td>
<td>83.3</td>
</tr>
<tr>
<td>Chan et al. 2015[81]</td>
<td>104</td>
<td>24</td>
<td>70.8</td>
<td>98.1</td>
</tr>
<tr>
<td>Kovacs et al. 2016[82]</td>
<td>60</td>
<td>15</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Saad et al. 2016[77]</td>
<td>114</td>
<td>62</td>
<td>63</td>
<td>82</td>
</tr>
<tr>
<td>Ruiz Hidalgo et al. 2016[83]</td>
<td>194</td>
<td>67</td>
<td>79.1</td>
<td>97.9</td>
</tr>
<tr>
<td>Ruiz Hidalgo et al. 2017[84]</td>
<td>44</td>
<td>23</td>
<td>61</td>
<td>75</td>
</tr>
<tr>
<td>Xu et al. 2017[85]</td>
<td>147</td>
<td>77</td>
<td>83.7</td>
<td>84.5</td>
</tr>
<tr>
<td>Ambrosio et al. 2017[86]</td>
<td>480</td>
<td>94</td>
<td>90.4</td>
<td>96</td>
</tr>
<tr>
<td>Sideroudi et al. 2017[87]</td>
<td>50</td>
<td>55</td>
<td>91.7</td>
<td>100</td>
</tr>
<tr>
<td>Francis et al. 2017[88]</td>
<td>253</td>
<td>62</td>
<td>90</td>
<td>91</td>
</tr>
<tr>
<td>Yousefi et al. 2018[89]</td>
<td>1970</td>
<td>796</td>
<td>88</td>
<td>14</td>
</tr>
<tr>
<td>Lopes et al. 2018[90]</td>
<td>2980</td>
<td>188</td>
<td>85.2</td>
<td>96.6</td>
</tr>
<tr>
<td>Steinberg et al. 2018[91]</td>
<td>105</td>
<td>50</td>
<td>63</td>
<td>83</td>
</tr>
<tr>
<td>Issarti et al. 2019[92]</td>
<td>312</td>
<td>90</td>
<td>97.8</td>
<td>95.6</td>
</tr>
<tr>
<td>Chandapura et al. 2019[93]</td>
<td>221</td>
<td>72</td>
<td>77.2</td>
<td>95.6</td>
</tr>
<tr>
<td>Xie et al. 2020[94]</td>
<td>1368</td>
<td>202</td>
<td>76.5</td>
<td>98.2</td>
</tr>
<tr>
<td>Kuo et al. 2020[95]</td>
<td>170</td>
<td>28</td>
<td>28.5</td>
<td>97.2</td>
</tr>
<tr>
<td>Shi et al. 2020[96]</td>
<td>55</td>
<td>33</td>
<td>98.5</td>
<td>94.7</td>
</tr>
<tr>
<td>Toprak et al. 2020[97]</td>
<td>66</td>
<td>50</td>
<td>94</td>
<td>98.5</td>
</tr>
<tr>
<td>Issarti et al. 2020[98]</td>
<td>304</td>
<td>117</td>
<td>85.2</td>
<td>70</td>
</tr>
<tr>
<td>Lavric et al. 2020[99]</td>
<td>1970</td>
<td>791</td>
<td>89.5</td>
<td>96</td>
</tr>
</tbody>
</table>

**Table 2.1:** Summary of the 26 published studies that included the use of machine learning for the detection of subclinical keratoconus
2.3.2.2 Corneal Imaging Data and Derived Parameters

An analysis of corneal images was used to detect subclinical keratoconus in 25 of the 26 papers. In the following sections, the use of quantitative measures derived from corneal imaging when used either in isolation or in combination within machine learning models is briefly discussed.

**Keratometry parameters**  Keratometric data was among the most popular parameters used in the literature, with 18 papers out of 26 incorporating keratometry as one of the parameters in their model [78, 76, 60, 82, 84, 96, 93, 101, 80, 99, 87, 77, 100, 81, 85, 89]. When looking at individual keratometric parameters derived via Fourier analysis for subclinical keratoconus detection, Sideroudi et al [87] achieved predictive accuracy of over 90% using higher order irregularities, asymmetry and regular astigmatism, mainly in the corneal periphery.

**Elevation parameters**  Sixteen papers incorporated elevation parameters in their analysis [78, 93, 98, 92, 79, 76, 82, 96, 86, 100, 81, 90, 80, 85, 89, 91]. Posterior corneal curvature quite consistently outperforms other parameters in discrimination of subclinical keratoconus [85, 76, 84, 78]. Its inclusion increases sensitivity of a support vector machine (SVM) from 75.2% to 92.0%, and precision from 57.4% to 78.8%, but has limited impact on specificity [76]. Posterior corneal curvature using the Pentacam was also found to be an important parameter for sensitivity using SVM, and much less so for specificity and Area Under the ROC Curve (AUC) [60]. Similarly, using the Galilei system, posterior asphericity asymmetry index (followed by corneal volume) was found to be the variable with most discriminatory power when differentiating normal from subclinical keratoconus [78]. Conversely, the random forest method was unable to discriminate subclinical keratoconus (VAE-NT), that is those fellow eyes of keratoconus patients with normal topography, from normal eyes using anterior surface topographic parameters and aberrometry [93].
Saad et al. (2016) showed that combining corneal wavefront data, thus using parameters obtained from the anterior corneal curvature, and Placido-derived indices led to a better discriminative ability between normal eyes and eyes with subclinical keratoconus (FFKC) over a Placido-based algorithm [77].

**Displacement parameters** Six papers used displacement parameters in their analysis [100, 79, 81, 82, 90, 101]. Several papers used the displacement of the thinnest point (DTP) from the geometric center of the cornea in their model [102, 100, 81]. Kovacs et al. (2016) used the vertical and horizontal decentration of the thinnest point and found them to be the best parameters to discriminate normal fellow eyes of keratoconus from control eyes using a neural network [82].

**Pachymetry parameters** Twenty papers used pachymetry data in their model [78, 60, 82, 96, 93, 89, 101, 98, 90, 85, 80, 92, 100, 81, 99, 84].

**Summary indices** Six papers used summary indices in their model [78, 60, 82, 103, 96, 93]. In a recent work, Shi et al. (2020) used 6 indices from the Pentacam along with keratometric, elevation, and pachymetry parameters derived from the Pentacam and Ultra High Resolution OCT (UHR-OCT) to create a neural network classifier for discriminating between normal and subclinical keratoconus eyes. They included 50 normal eyes, 38 eyes with keratoconus and 33 eyes with subclinical keratoconus. They achieved 98.5% sensitivity and 94.7%, however the result requires further validation due to the small number of eyes in this group (Shi et al., 2020). Furthermore, the authors did not include a comparison between existing detection metrics such as BAD_D.

**Heat Maps** Two papers used heat maps in their detection model [104, 95]. An emerging trend in the diagnosis of subclinical keratoconus is the use of coloured maps generated from quantitative analysis of raw images. Modalities such as Scheimpflug and AS-OCT capture images at various corneal meridians and subsequently use this data to derive the heat maps that facilitate visual interpretation of
the data, although there is extrapolation of the data in areas between the imaged meridians. For example, the Pentacam can translate the raw images into several types of colour heat maps (e.g. axial curvature, posterior/anterior elevation, regional pachymetry), all based on the same original tomography dataset. Prediction models applied to images often use convolutional neural networks (CNNs) and studies applying these methods will be discussed in more detail in the machine learning section. To the best of the author’s knowledge, no system has attempted to use raw Scheimpflug or AS-OCT images directly when predicting subclinical keratoconus.

2.3.2.3 Biomechanical data

Three papers incorporated biomechanical data in their analysis [86, 88, 91]. Numerous publications describe the application of machine learning to analyse biomechanical data, however only the few which validated their results were included in this review. Ambrosio et al. combined Pentacam and Corvis ST data to create the Tomographic and Biomechanical Index (TBI) [86] and this was followed up with a validation study [91]. Francis et al. used biomechanical data from the Corvis ST device when diagnosing keratoconus and achieved very high sensitivity (99.5%) and specificity (100%), however when validating their model, they appear to have only discriminated between two groups: combined subclinical keratoconus and keratoconus eyes from normal eyes, which presents an easier problem than including a distinction between normal and subclinical keratoconus eyes [88].

2.3.2.4 Demographic Risk Factors

A small number of studies chose to include demographic data, such as age or gender, in their model [60, 90, 96, 78]. Cao et al. (2014) showed that gender was an important parameter in a minimum set that achieved the highest AUC using the random forest method although their dataset was small (49 subclinical keratoconus, 39 control eyes) [60]. Notably ethnicity, a major factor associated with disease prevalence, risk of progression, and disease severity, such as hydrops being much more prevalent in Asian and Black populations [105], was not included in any model, although
some reported studying single ethnicities [104]. Ethnicity as a parameter should be considered by future investigators. No papers included recognised risk factors such as atopy and eye rubbing as a model parameter.

2.3.3 Machine Learning Techniques

In most cases, researchers opted to use combinations of parameters and indices to diagnose subclinical keratoconus. This section is subdivided according to the machine learning techniques that were applied. Figure 2.1 summarises these techniques.

Figure 2.1: Organisational diagram of relevant machine learning algorithms reported to be used for the detection of subclinical keratoconus. There are many other algorithms but discussion of these is beyond the scope of the review. Only the methods found in the results were included.

2.3.3.1 Neural Networks

Neural networks consist of a series of interconnected layers of neurons and are thus loosely modelled on the structure within the human brain. Each neuron computes a nonlinear function of its inputs and the network is trained until the output aligns optimally with the ground truth labels. Kovacs et al. (2016) used a combination of 15 keratometric, pachymetric, and elevation parameters in a neural network classifier to discriminate healthy corneas from fellow eyes of patients with unilateral
keratoconus [82]. The patient data included 60 normal eyes from 30 patients, 60 bilateral keratoconus eyes for 30 patients and 15 normal eyes from patients presenting unilateral keratoconus. When classifying the normal eyes of patients with unilateral keratoconus with clinical grading as a reference, they achieved 90% sensitivity and 90% specificity. They took the novel approach of training on both eyes of the patients, which allowed them to incorporate the effect of any inter-eye asymmetry when diagnosing unilateral keratoconus. However, using both patient eyes to differentiate between unilateral keratoconus and normal eyes is similar to differentiating bilateral keratoconus from normal eyes because the diagnosis in the diseased eye is not in doubt. A significantly harder problem, and the main focus of this review, is to identify subclinical keratoconus (if it affects both eyes, or without knowledge of the disease status of the fellow eye in unilateral keratoconus). Shi et al. combined keratometric, elevation and pachymetry parameters derived from Pentacam images and UHR-OCT to create a neural network classifier for discriminating normal from subclinical keratoconus eyes [96]. Using Pentacam elevation and pachymetry maps within a hybrid neural network model [92] demonstrated superiority over other common diagnostic indices such as BAD-D and TKC although, interestingly, it was similarly discriminatory to such indices for mild and moderate keratoconus against normal [92]. A feedforward neural network trained on Pentacam anterior and posterior curvature parameters from normal, subclinical and a spectrum of keratoconus patients performed similarly to BAD_D in detecting confirmed keratoconus, and was also better at identifying subclinical keratoconus [98].

2.3.3.2 Convolutional Neural Networks

When images are used for analysis, CNNs are often employed as the machine learning algorithm due to their ability to make inferences from two or three dimensional data structures. For example, Yi Xie et al. (2020) used data from 1368 normal eyes, 202 eyes with early keratoconus, 389 eyes with more advanced keratoconus and 369 eyes with subclinical (suspected) keratoconus to develop an automatic classifier. They achieved 76.5% sensitivity and 98.2% specificity when classifying subclinical
keratoconus [104]. However, the heat maps they used were produced by Pentacam, which means that the technique may not be transferable to other systems or even future iterations of the Pentacam software. Kuo et al. (2020) included 150 normal, 170 keratoconus and 28 subclinical eyes in their study and used the TMS-4 topography system to produce corneal heat maps and trained 3 different CNN architectures (VGG16, InceptionV3 and ResNet152). When attempting to identify the 28 subclinical keratoconus eyes, they applied the VGG16 model and achieved ‘barely satisfactory’ results with an accuracy of 28.5% when a threshold of 50% was applied [95]. From these results, it can be seen that subclinical keratoconus cannot yet be detected with high sensitivity using heat map images.

2.3.3.3 Decision Trees

Decision trees use a branched tree structure to perform classification. The classification of data in a decision tree is similar to a flow chart, where a binary decision at each node in the tree determines which branch to take next. Starting from the root, the classification (e.g. keratoconus, subclinical keratoconus, normal) is found by following each branching to its terminal node. Smadja et al. (2013) used a decision tree for classifying between normal, keratoconus and subclinical (FFKC) keratoconus eyes. They enrolled 177 normal eyes, 148 keratoconus eyes and 47 subclinical eyes. They used 55 parameters (including curvature, elevation, corneal wavefront, corneal power, pachymetry, age) collected from the Galilei dual Scheimpflug camera and were able to achieve 93.6% sensitivity and 97.2% specificity when classifying subclinical from normal [78]. Cao et al. (2014) also evaluated a decision tree algorithm for classifying subclinical keratoconus but achieved lower sensitivity (82%) and specificity (78%) [60]. They attribute their inferior performance to the fact that Smadja et al. had used additional machine-specific indices that they did not have access to.
2.3.3.4 Random Forests

Random forests combine several decision trees into a single model. Lopes et al. (2018) compared this method along with other methods (naive bayes, neural networks, SVM, discriminant analysis) by training models on 71 post-LASIK eyes with ectasia, 298 post-LASIK eyes without ectasia and 183 eyes with keratoconus. They included keratometry, pachymetry, elevation, and various Pentacam indices. They validated their models on an external dataset containing 298 normal eyes (stable LASIK), 188 keratoconus eyes [very asymmetric ectasia-ectatic (VAE-E)], and 188 subclinical eyes [very asymmetric ectasia-normal topography (VAE-NT)]. The latter two groups were collected from the same set of patients. They found that the random forest model performed best when detecting subclinical eyes with 85.2% sensitivity. This accuracy is lower than other comparable studies, which is probably due to their inclusion of external validation rather than an inferior model [90]. The authors also note the fact their model is classifying amongst 3 groups whereas other related studies (such as [78]) are only classifying between 2 groups (e.g. subclinical vs normal). This is an important distinction and will be expanded upon in the discussion.

2.3.3.5 Discriminant Analysis

Discriminant analysis uses a linear combination of variables that optimally separates two or more classes of data. Xu et al. (2017) used this method to classify eyes as either normal, subclinical keratoconus or keratoconus [85]. They included 147 normal eyes, 139 eyes with keratoconus, and 77 with subclinical keratoconus in their training set and verified on a separate set of 97 normal and 49 subclinical keratoconus eyes. They applied the Zernike fitting method to corneal pachymetry and elevation data derived from the Pentacam and were able to achieve 92.8% AUC when discriminating subclinical keratoconus.

Although the original paper was not directly included in the results due to lack of validation, Saad et al (2010) [102] created a discriminant analysis model using a dataset consisting of 72 normal, 40 subclinical (FFKC), and 31 keratoconus eyes which was later validated by Saad et al. (2012) [100] and Chan et al. (2015) [81].
Saad et al. (2012) [100] reported sensitivity of 92% and specificity of 96%, making it the best performing model after validation and this is discussed further in section 2.3.4.

A later piece of research by Saad et al. (2016) also used discriminant analysis to classify eyes as either subclinical keratoconus (FFKC) or normal. They used a combination of wave-front aberrometry and placido disc indices in their model with a total of 8 parameters. The model was trained on 114 normal and 62 subclinical eyes and validated on 93 normal and 82 subclinical eyes. Using training data only, their model achieved 89% sensitivity and 92% specificity, but when applied to the validation set, the accuracy dropped significantly to 63% sensitivity and 82% specificity [77]. This highlights the need for external validation when reporting the performance of detection algorithms.

2.3.3.6 Support Vector Machines

Support vector machines (SVM) translate data into a higher dimensional space where a dividing line (known as a hyperplane) separates the data such that the distance between the hyperplane and any given data point is maximised. When eight different machine learning algorithms were compared for classifying subclinical keratoconus on the same dataset, SVM achieved the highest sensitivity (94%) [60]. Arbelaez et al. (2012) achieved even higher sensitivities using SVM on a large dataset of 1259 normal eyes and 426 with subclinical keratoconus [76]. Two hundred eyes from each group were used for training with the remainder for testing, achieving 92% sensitivity and 97.7% sensitivity. Ruiz Hidalgo et al. (2017) used 25 topographic/tomographic Pentacam derived parameters to verify their SVM model. They included 131 patients in their study and provided results for two classifications from separate hospitals: Antwerp University Hospital and Fondation Rothschild, Paris. When classifying between four groups (keratoconus, subclinical, normal, post refractive surgery), their sensitivity for subclinical keratoconus detection was 61% relative to the Antwerp University Hospital classification and 100% relative to the Rothschild classification. This was a comprehensive validation study which
compared multiple methods against two subjective reference standards [84]. Unfortunately, only a small number of subclinical keratoconus cases (20) were included in this study, and a larger study is required to fully verify the result.

2.3.3.7 Logistic Regression

Logistic regression is commonly used to combine multiple outputs to predict a binary outcome. A sigmoid function is fitted to the data, returning a probability value between 0 and 1, which is then thresholded for classification. Three studies used this technique exclusively when classifying subclinical keratoconus [88, 87, 101]. Sideroudi et al. (2017) used logistic regression to explore the diagnostic capacity of Fourier-derived posterior keratometry parameters (spherical component, regular astigmatism, asymmetry, irregular astigmatism) extracted from Pentacam Scheimpflug images [87]. They included 50 normal eyes, 80 eyes with keratoconus, 55 with subclinical keratoconus (defined as a clinically normal eye with abnormal topography, where the fellow eye has advanced KC), and validated their model on 30% of the dataset. Their model attained 91.7% sensitivity and 100% specificity when classifying between subclinical keratoconus and normal eyes. Although these results are among the best reported, the study has yet to be validated on an external dataset. Other papers implemented logistic regression as part of a wider comparison of machine learning algorithms [86, 60].

2.3.3.8 Comparative Studies

Only a few papers have applied multiple machine learning algorithms to the same dataset. Cao et al. (2014) tested 8 machine learning algorithms on the same dataset of 39 normal control eyes and 49 eyes with subclinical keratoconus [60]. They used 11 parameters: age, gender and 9 corneal parameters from Pentacam tomography, and found that random forest, SVM, and K-nearest neighbours had the best performance. Random forests had the highest AUC with 0.97, SVM had the highest sensitivity (94%), and K-nearest neighbours had the best specificity (90%). Although they did verify their results with 10-fold cross validation, repeating the analysis on
a larger dataset would be informative. Ambrosio et al. (2017) also performed an analysis across algorithms including logistic regression, SVMs and random forests to classify between four groups: normal, keratoconus, VAE-E, and subclinical keratoconus [very asymmetric ectasia-normal (VAE-N)] [86]. They used both Scheimpflug tomography and biomechanical data and included 480 normal eyes, 204 eyes with keratoconus, 72 eyes classed as VAE-E and 94 subclinical keratoconus eyes. When considering subclinical keratoconus, the random forest model performed the best, with 90.4% sensitivity and 96% specificity. They named their final model the Tomography and Biomechanical Index (TBI) and validated it by leave-one out cross-validation resulting in as many models as there were subjects (850). A new model is built for all cases, excluding one case in which the model is tested. Lopes et al. (2018) also performed a comparative analysis and found that random forests performed best when trying to classify 3 groups of eyes (including subclinical eyes) [90]. Lavric et al. (2020) provided the largest of comparative studies for detecting subclinical keratoconus [99]. They included 1970 normal, 390 eyes with keratoconus and 791 subclinical (FFKC) keratoconus eyes in their study and used keratometry, pachymetry, and aberrometric data from the CASIA AS-OCT system in their analysis across 25 different machine learning algorithms. When they classified the 3 groups simultaneously they found the most accurate method was SVM, which attained 89.5% sensitivity for the detection of subclinical keratoconus, and the results were validated using 10-fold cross validation. Some limitations of this study include the use of the CASIA ESI index for classification of the severity of keratoconus, which may not agree with clinical diagnosis, and the fact that the analysed parameters are closely tied to the CASIA device, which limits its generalisability to other systems.

### 2.3.3.9 Unsupervised Learning

Unsupervised learning represents a novel approach in the detection of subclinical keratoconus. These methods attempt to identify groups of similar eyes without using pre-labelled data. Yousefi et al. (2018) used this technique to cluster a cohort of 3156 eyes categorised according to the ESI index as either normal, keratoconus
and subclinical (FFKC) keratoconus. They included 420 topography, elevation and pachymetry parameters and the algorithm produced four clusters of eyes with similar characteristics. When comparing their results to a reference standard (ESI index), the model did not create a distinct grouping that separated the subclinical eyes from other eyes, with calculated sensitivity of 88% and specificity of 14%, suggesting poor correlation compared to ESI alone. Furthermore, they did not compare their results to clinically labelled data [89].

2.3.4 Validation

Whilst most studies performed internal validation by splitting their original dataset into training and test sets, a number of replication papers which validated a published model on a new dataset were identified. Ruiz Hidalgo et al. (2017) [84] verified their SVM technique, which was presented in 2016 [83]. The authors found that when using the Antwerp University Hospital classification, there was an approximate 18% decrease in sensitivity whereas when using the Rothschild classification, there was an approximate 21% increase in sensitivity. These discrepancies highlight the issues associated with subjective classification. Furthermore, when multiple groups were included in the analysis, such as normal, keratoconus, subclinical keratoconus, and post-refractive surgery eyes, it was notable that the accuracy decreased from 93.1% in discriminating normal from FFKC, to 88.8%. However, this paper represented the most comprehensive methodology, as the authors not only verified their results on a new sample population with multiple target classes, but also compared their results against other methods and included two subjective reference standards.

Buhren et al. (2014) validated their model defined in 2010 [106]. When comparing their discriminant function, derived from anterior and posterior corneal surface wavefront data, they reported an approximate 22% decrease in sensitivity and an approximate 9% decrease in specificity [80]. This decrease is likely due to overfitting in the original model. Saad et al. (2012) [100] and Chan et al. (2015) [81] both validated the same discriminant analysis model presented by Saad et al (2010) [102]. Saad et al. reported sensitivity of 92% and specificity of 96%, roughly in
line with their previous study, which indicates that their method is reliable and does not suffer from overfitting [100]. Chan et al. [81] validated the original model in patients from a different ethnic background (Asian). They reported a 21% decrease in sensitivity which they attributed to overfitting in the original study however their specificity was almost equivalent. Steinberg et al. (2018) [91] validated the work presented in Ambrosio et al (2017) [86]. They reported a 27% decrease in sensitivity and a 13% decrease in specificity when applying the same cutoffs.

2.4 Discussion

2.4.1 Data Types for Detection of Subclinical Keratoconus

The data used for building algorithms for detecting subclinical keratoconus can be numeric parameters or heat map images derived from these parameters. The parameters are obtained from a number of different imaging systems and devices, and then incorporated in different combinations to build a classification system. Inevitably, individual systems produce parameters that may not be comparable across devices and, for proprietary reasons, the raw data may not be available. Comparison or replication across systems can therefore be difficult. Heat maps provide a helpful visual representation of corneal elevation, pachymetry or curvature for interpreting results. However, heat maps require interpolation or extrapolation of data which may introduce inaccuracies when included in the model. No studies were found that have analysed pixel-level corneal Scheimpflug imaging data, likely due to restricted access on commercial machines and the difficulty of extracting this data from a large number of scans in an automated manner.

Additionally, many studies do not include patient demographic and associated disease, such as age, gender, ethnicity and atopic diseases. Incorporating this data into these models may help define the population to which an algorithm applies, particularly as there are phenotypic indices that an algorithm can identify from images that humans cannot identify by manual inspection [107].
2.4.2 Machine Learning

Published studies typically involve univariate or multivariable analysis. For univariate studies, ROC analysis was performed to quantify the diagnostic ability of each parameter. However, as none of the papers identified performed an out-of-sample validation, they were excluded. For multivariable studies, machine learning is used to create a detection model from multiple parameters. These algorithms have already demonstrated comparable performance to experienced ophthalmologists in the identification of retinopathy of prematurity [108] and retinal disease progression [109]. Machine learning-based research to detect subclinical keratoconus has primarily focused on supervised learning techniques, such as decision trees, SVM, logistic regression, discriminant analysis, neural networks, and CNNs. Logistic regression may be superior to neural networks when parameters from a single imaging modality are considered [60, 96] with potentially a greater role for neural networks when a large number of potentially interacting parameters are combined, such as in the case of multiple imaging modalities [96]. Unsupervised learning has also been evaluated for detecting subclinical keratoconus, however it relies on identifying patterns in large amounts of data and thus may not translate to a different dataset of a different size and with different properties. In addition, with the exception of Yousefi et al (2018)[89], none of the papers provided access to their algorithms’ source code or descriptions of hyper-parameters, which makes it difficult to reproduce and validate the results with external datasets.

2.4.3 Validation

Studies that did not include a validation arm were excluded, and the vast majority of initially identified studies did not appropriately validate their results. Validating results on a dataset separate from the training set is essential for determining the generalisability of an automatic classifier to other datasets. Apart from Saad et al (2012) [110] and Hidalgo et al (2017) [84], studies attempting to validate previous methods reported significant decreases in sensitivity and specificity compared to their initial results. This highlights that, even when techniques like cross-validation
are used, validating with an independent out-of-sample dataset is the most effective method. Ideally, this dataset should be larger and more representative of the general population.

2.4.4 Strengths and Limitations

The primary strength of this review is its focus on subclinical keratoconus detection, an important and useful clinical question, rather than clinically obvious disease. The review complements recent clinical trials in the role CXL treatment has in preventing keratoconus progression in children and young adults [111, 35]. However, several limitations stem from the search method and inclusion criteria. Articles which did not include the relevant key terms or were not appropriately indexed by the literature databases may have been missed. Furthermore, because of limitations such as discrepancies in the definition of subclinical keratoconus across publications, some studies may have not been included. Finally, articles which did not perform validation on their results were excluded, further reducing the pool.

2.4.5 Challenges and Future Directions

Study Design None of the reported studies evaluated the performance of their method against masked observers. The initial classification is often made by taking account of the fellow eye with keratoconus as a factor in the decision making process, whereas the algorithm does not have this information. It would be interesting to design a study where, having already decided on the ground truth diagnosis, a new clinician is asked to evaluate the eye using exactly the same information as the algorithm (i.e. only the images or parameters). This situation is closest to real life screening where a prospective patient (without a history of keratoconus in either eye) is examined for risk of keratoconus.

Subclinical keratoconus is, by definition, the least affected eye of highly asymmetric keratoconus. An assumption is that any parameters of subclinical disease that differ from the values for normal corneas are the result of keratoconus. However, it has not been demonstrated prospectively that all of the eyes in such a cohort
will progress to the clinical disease state. Although true unilateral keratoconus is thought not to exist [29], this has not been proven, and it is possible that some eyes with subclinical keratoconus are not at risk of progression and that some of the abnormal parameters in this group are not the result of keratoconus. It would be valuable to conduct a prospective study where eyes which do not develop clinical keratoconus over time are used as lower risk examples.

**Study Size and Statistical Power** Study size is crucial when developing a dependable detection system, particularly as the accuracy of machine learning models is closely tied to the quantity of training data. Only two studies of eyes with subclinical keratoconus included more than 500 eyes [99, 89]. Importantly, none of the papers carried out a priori power calculations to determine the size of the cohort to be studied.

**Case Definition, Gold Standard, and Ground Truth** Precise comparisons between the results of these publications is problematic due to the ambiguous definition of early keratoconus and the absence of a ‘gold standard’ examination technique. The most common definition of subclinical keratoconus is an eye with topographic findings that are at least suspicious of keratoconus and with confirmed keratoconus in the fellow eye. FFKC is most usually defined as an eye that has both normal topography and slit-lamp examination but with keratoconus in the fellow eye [69]. With this differentiation, subclinical keratoconus will be easier to detect than FFKC and studies using the former definition are likely to produce more accurate results because the problem becomes easier to solve. The problems of making statistical comparisons in the absence of a gold standard has been discussed extensively [112]. The authors suggest that latent class analysis, composite reference standards or expert panel analysis may be appropriate in these circumstances.

Even if a precise definition of early subclinical keratoconus were established, the absence of ground truth data is relevant when evaluating the precision of the data acquisition. For example, measurements of keratoconus taken by different operators, or repeated on different days, may lead to a variation in the results. Flynn
et al. found that keratometric measurements from Scheimpflug images (Pentacam) were more reproducible in early keratoconus (Mean central K \leq 53D) compared to more advanced keratoconus (Mean central K > 53D), although a cohort with subclinical keratoconus was not included [63]. In contrast, Yang et al. (2019) found that biomechanical parameters (Corvis ST) had an acceptable repeatability in both normal and keratoconus eyes [113].

A further issue identified when comparing studies was variation in the number of groups that were classified. The studies often started with multiple groups (usually three, e.g. FFKC, keratoconus, normal), but then 21 papers chose to report their accuracy results from a model trained to classify between just two groups (e.g. FFKC and normal), whilst 5 papers reported results for classifying between all groups. Classifying all groups is a more realistic clinical scenario but it presents a more challenging problem because features of the different groups can overlap. Full details of the number of groups associated with the accuracy results are presented in Multimedia Appendix 2 from the published text [8].

New Avenues of Research  Despite the growing literature on the diagnosis of subclinical keratoconus, there is a need for further fundamental research, particularly for analysis based on the raw pixel values rather than only derived parameters. Furthermore, a multimodal solution could be developed by combining these raw images with other parameters such as biomechanical, demographic and genetic data. Demographic data such as age, gender, ethnicity and allergic eye disease are known risk factors for progressive keratoconus, and a family history of keratoconus is also a likely risk factor that should also be included in diagnostic algorithms. Environmental risk factors including eye rubbing have been associated with keratoconus progression, although eye rubbing is difficult to quantify. A genetic predisposition to keratoconus is supported by heritability studies in twins, linkage analysis in families and population level genome-wide association studies [114, 115]. From these studies, genetic risk scores have been derived which could, in the future, be included in machine learning models for the detection of subclinical keratoconus. Ideally, a prospective study should be performed in a large cohort of young (< 30 years
of age) patients with subclinical keratoconus to monitor for disease progression. Training should be conducted on large datasets with the explicit aim of detecting subclinical keratoconus and the resulting model should be externally validated on a new dataset. Important discriminatory parameters to be considered include posterior elevation and corneal volume. Finally, a range of machine learning techniques should be applied to the same dataset along with detailed comparison statistics.

### 2.4.6 Conclusion

This chapter represents the most comprehensive review to date of machine learning algorithms for the detection of subclinical keratoconus. Early detection of keratoconus to enable treatment is a public health priority, and the utilisation of machine learning algorithms has the potential to make the diagnostic process more efficient and widely available. The relevant publications were summarised in terms of their input data and choice of algorithm, ensuring all studies performed appropriate validation. Whilst comparison between methods is challenging, the best performing externally validated algorithm used discriminant analysis applied to a combination of numeric corneal parameters. The difficulty of obtaining accurate datasets for training machine learning algorithms and the need for a consistent, objective and agreed definition of subclinical keratoconus was identified. No publications were found that used raw images as input and as such, new avenues of research have been identified that combine raw image data with biomechanical, demographic and genomic data. Chapter 3 explores the idea of working on raw images further. In addition, defining disease progression and modelling its course is an area which may benefit from further research. This up-to-date review is important for all researchers in this field, for those in clinical practice and public health, as well as for policy makers to understand the current state of the research and provide guidance for future health service planning.
Chapter 3

Early versus Late Stage Keratoconus Classification Using Deep Learning
3.1 Introduction

As analysed and discussed in chapter 2, machine learning has, in recent years, been applied to keratoconus detection with encouraging results. In particular, deep learning, a subfield of machine learning has shown promise in classifying disease from images. Many of the existing deep learning techniques rely on the use of derived data such as heat maps. Furthermore, none of these algorithms have achieved satisfactory accuracy when detecting subclinical disease. Using raw image data has been applied in some research when classifying between different corneal pathologies such as Fuchs, Dry Eye Syndrome (DES), and keratoconus [116]. Whilst the ultimate aim should be to detect subclinical keratoconus from normal patients, constraints on time and data availability meant that a more straightforward machine learning task had to be tackled. Within this chapter, a simplified version of keratoconus staging was addressed by attempting to classify early vs late stage keratoconus using corneal images from the MS-39 device which outputs N radial slices of the cornea per scan. Although this basic problem does not serve a clinical purpose, it is pursued as a proof of concept that raw images can be used to detect differences in corneal anatomy relevant to keratoconus. Future work could extend the problem to multi-stage classification and ultimately subclinical keratoconus detection.

Two approaches to this task are attempted. The first approach (Approach 1) involves using the raw images directly for training a CNN. Potential improvements to this approach may be gained by using an image without artefacts and other parts of the corneal anatomy. Therefore, the second approach (Approach 2) initially segments the cornea from the raw images before using the segmented images to train the same CNN.

3.2 Data

The dataset consists of 368 scans corresponding to unique eyes from 348 patients at the Early Keratoconus Clinic (EKC) within Moorfields Eye Hospital (MEH) who visited between June 2020 and December 2022. Patients were split into two groups:
<table>
<thead>
<tr>
<th>Type</th>
<th>Slices / scan</th>
<th>Resolution</th>
<th>Stage 1 scans</th>
<th>Stage 4 scans</th>
<th>Patients</th>
<th>Ap. 1</th>
<th>Ap. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>12</td>
<td>1800x600</td>
<td>133</td>
<td>133</td>
<td>255</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>B</td>
<td>25</td>
<td>1800x1024</td>
<td>51</td>
<td>51</td>
<td>93</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Table 3.1:** Table summarising the dataset in terms of scan type. Each scan corresponds to a unique eye and comprises of 12 (type A) or 25 (type B) radial slices. Ap. = Approach.

stage 1 (Front K2 \(<48\)D and pachymetry \(<530\)µm) and stage 4 (Front K2 \(>55\)D) yielding a balanced dataset of 184 scans for both stage 1 and stage 4. The decision to focus on classifying stage 1 vs stage 4 (as opposed to other combinations of stages) was taken to allow for a genuine separation between patients’ eye topography and thus make this exploratory work more achievable in its initial stage. Images were captured by the MS-39 AS-OCT device. The device is capable of producing two main types of scans (referred to as type A and type B going forward) with different resolutions and slices per scan (see table 3.1). Ideally, all scans would be of the same type but given the overall lack of scans, we chose to analyse as many as possible, leading to two different sets of scans. Table 3.1 summarises the distribution of these scan types within the dataset. There were a total of 3192 (12x266) type A images and 2550 (25x102) type B images.

### 3.2.1 Pre-processing

All images were cropped by selecting the top third only (600 pixels) because clinical diagnosis is usually performed by examining both the curvature and thickness of the cornea (ignoring the deeper part of the anterior chamber). See 3.1 for an illustration of the cropped region used for training the CNN. The images were also normalised between 0 and 1 by dividing each pixel value by 255. These steps were performed for both approaches.

The second approach applied segmentation to try and remove image artefacts. In order to train a segmentation algorithm, a dataset of annotated images is required. Manual annotation of corneal images is a time consuming process and extensive la-
belling from clinical collaborators was not available at the time of writing. However, with the help of fellow researchers (LK), it was possible to annotate 120 slices (from 40 scans) from both stage 1 and stage 4. Annotation was only available for image type B. No further pre-processing of images was performed.

Figure 3.1: Example AS-OCT image slice from the MS-39 device for image type B. The original dimensions were 1800x1024 but the upper part of the images were cropped to 600x1024 (as indicated by the red box) when training and validating the CNN.

3.3 Methods

3.3.1 Classification Model

Both classification approaches used the same neural network architecture. A custom neural network was used because initial exploration yielded better results than using existing architectures such as ResNet or VGGNet. The structure of this network can be seen in figure 3.2. The input layer requires the image to be transformed to 224x224. After this there were two convolution blocks with filter size 10 and depths 32,64 respectively. The convolution layers used ReLu activation and were
separated by 5x5 max-pooling layers with stride five. The network was flattened and then connected to a fully connected layer of size 1024. This was followed by ReLU activation and a dropout layer (probability 0.5). The final layer was a softmax classification layer for the two classes.

Figure 3.2: Custom CNN architecture for classification of stage 1 vs stage 4 keratoconus. Generated using Net2Viz [117]

3.3.2 Segmentation Model

Approach 2 requires a segmentation model. Given the relatively small size of the dataset, it was decided to use a patch based segmentation technique [118, 119]. Patch based segmentation is a form of augmentation where patches of size NxN and their labels are extracted from the dataset and fed into the network in batches. This gives the network many more images to train with. A patch size of 512x512 was used along with a batch size of 16. A basic U-Net architecture was used for training (courtesy of BL) which can be seen in figure 3.3.

3.3.3 Approach 1: Classification Only

This first approach aims to classify early vs late stage keratoconus using the raw images as input to a basic CNN. The top section of figure 3.4 illustrates the training process for approach 1 schematically. As noted in the explanation of the data, there are two types of images and using both within the same model was deemed to be problematic. Therefore, in order to make full use of the dataset, a separate model was trained for image type A and image type B using an identical method.
Figure 3.3: U-Net architecture used for training segmentation algorithm. Generated using Net2Viz [117]

Figure 3.4: Diagram illustrating the training process for the two approaches: Approach 1 and Approach 2. Approach 1 applies a CNN to the raw image whereas Approach 2 first applies a U-Net to identify the cornea and blacks out the rest of the image.

As a supplementary analysis to verify that the network was genuinely learning from the corneal anatomy and not some latent information common to the image acquisition (e.g. lighting, artefacts), a sanity experiment was performed where only a 100x100 patch was used as input to the network. The expectation from this experiment was that classification accuracy should be close to 50%.
3.3.3.1 Training and Validation

The data was divided into training, test and validation sets using a 70:15:15 split. The model were therefore trained and tested using 85% of the dataset and the validation set (15%) was kept separate for independent evaluation of the model. The network was trained for 50 epochs using binary cross entropy loss as the loss function:

\[ \text{Loss} = y \ln(p) + (1 - y) \ln(1 - p) \]

where \( y \) is a binary indicator (0 or 1) for the ground truth label for a given image, \( p \) is the predicted probability for a given image and \( \ln \) is the natural logarithm.

Early stopping with patience of 20 epochs was used to terminate training when there had been no further improvement. Adam optimiser was selected as the optimiser. For the hyperparameters, a learning rate of 5e-5, weight decay of 1e-4 to reduce overfitting and a batch size of 64 were applied.

3.3.4 Approach 2: Segmentation Followed by Classification

Although the upper third of the image was used for training approach 1 with the hope of focusing on the cornea only, figure 3.5 represents an example of an image that contains artefacts such as a vertical streak and parts of the eyelid. Therefore, the second approach improves the method from Approach 1 by performing segmentation of the raw images before classification. The lower section of figure 3.4 illustrates this process. It first performs segmentation using U-Net and then combines the mask with the original image to isolate the cornea without artefacts and subsequently trains the CNN as before. It should be noted that only the type B images were available for annotation and thus the model was trained on image type B only.

3.3.4.1 Training and Validation

For the segmentation step, two separate models corresponding to the two different stages were trained because initial results found this to give optimal accuracy.
Figure 3.5: Example MS39 image exhibiting artefacts (vertical streak) and unwanted anatomy (eyelids) in the upper third of the image.

Figure 3.6: Example original image (a) and the same image with its mask applied (b) for training approach 2.
The dataset was split with a ratio of 80:20 (training:test) and training was iterated 40,000 times. Adam optimizer was applied with a learning rate of 5e-5 and binary cross entropy was the choice of loss function.

Once the two segmentation models (separate for stage 1 and stage 4) were trained, they were applied to the corresponding raw images in order to generate masks. The masks were then applied to the same raw images to isolate the cornea. An example of a raw image before and after applying the mask can be seen in figures 3.6(a) and 3.6(b). The masked images were then split into training, test and validation as before with a ratio of 70:15:15 and the training process was repeated using the same optimiser, loss function and hyperparameters as section 3.3.3.1.

3.3.5 Software

All experiments were performed on a single NVIDIA GeForce RTX 3090 GPU and the classification training process took approximately 15 minutes. Software development was conducted within Python (version 3.8.2) and Keras (version 2.4.3) was used as the deep learning framework.

3.4 Results

3.4.1 Approach 1: Classification Only

Figures 3.7 and 3.8 shows the learning curves (both accuracy and loss) for the main classification experiment. There are different results for image type A and B only because it was problematic to combine the datasets. Regardless of image type, as expected, the training set loss is lower than the testing set loss which is a common phenomenon in machine learning. On the validation set for image type A, the final accuracy, sensitivity and specificity was 94.4%, 92.9% and 95.8% respectively and for image type B it was 93.5%, 94.0% and 93.0% respectively. Whilst it may initially be surprising that the image type A model had higher validation accuracy despite having higher loss, this can be explained by the fact that the image type A dataset
had nearly three times the number of unique eyes to learn from (as seen in Table 3.1), which could have enabled the model to better generalise and achieve higher accuracy. Figure 3.9 shows the confusion matrix for the validation sets using image type A and type B.

As a sanity experiment, figure 3.10 shows the learning curve (both accuracy and loss) where only a 100x100 patch was used for training. As expected, the model severely overfits which can be seen by the training accuracy being far higher than the test accuracy. Validation accuracy in this case only reached 55% which confirms that the images cannot be classified from small patches and thus corneal anatomy is fundamental in the classification process.
Figure 3.8: Approach 1 (image type B): Accuracy and loss curves for the stage 1 vs stage 4 classification problem. The blue line corresponds to the training set and the orange line corresponds to the test set.

Figure 3.9: Approach 1: Confusion matrix showing the classification counts of the CNN when classifying stage 1 vs stage 4 keratoconus for image type A (left) and image type B (right).
3.4.2 Approach 2: Segmentation Followed by Classification

3.4.2.1 Segmentation

When considering the separate segmentation models trained from stage 1 images and stage 4 images, the final dice scores were comparable at 94.8% and 94.5% respectively. Example inputs and outputs of the algorithm can be seen in figure 3.11. Although a few minor artefacts remain, it is clear that the model has managed to segment the images well. The training loss for each segmentation model can be seen in figure 3.12 and figure 3.13. Both loss curves dropped rapidly at first and then decreased gradually before no further improvement was attainable. The final loss for the stage 4 model was considerably higher than the stage 1 model which may be due to the fact that the shape of stage 4 corneas was more variable. However, the comparable dice scores indicate that the stage 4 model was still able to successfully perform segmentation, despite the higher loss.
Figure 3.11: Example input images (a and c) and predicted masks (b and d) for stage 1 and stage 4 images. The corneal segmentation is mostly correct but some artefacts remain. This could potentially be addressed with either more training data or post-processing. All images belong to type B.

3.4.2.2 Classification

When the classification model was retrained by applying the segmentation masks to the images during the training process, the new validation accuracy, sensitivity and specificity was 96.5%, 95.0% and 98.0% respectively. The confusion matrix for this experiment can be seen in figure 3.15. The small improvement in classification accuracy can be attributed to the improved quality (without artefacts) of the segmented images which allows the network to focus on just the areas of interest.
Figure 3.12: Training loss curve for the patch based segmentation algorithm applied to stage 1 images (image type B).

Figure 3.13: Training loss curve for the patch based segmentation algorithm applied to stage 4 images (image type B).
Figure 3.14: Approach 2: Accuracy and loss curves for the stage 1 vs stage 4 classification problem with pre-segmentation. The blue line corresponds to the training set and the orange line corresponds to the test set.

Figure 3.15: Approach 2 (image type B): Confusion matrix showing the classification accuracy of the CNN when classifying stage 1 vs stage 4 keratoconus after applying the segmented masks.
3.5 Discussion

In this chapter it has been shown that raw AS-OCT images can be used for the classification of early versus late-stage keratoconus with promising results, although the results remain to be externally validated. Approach 1 did not perform any segmentation of the raw images and achieved 94.4% and 93.5% accuracy for image type A and image type B respectively.

Approach 2 pre-segmented the raw images before completely retraining the same network from Approach 1. This technique improved the classification accuracy slightly from 93.5% to 96.5%. Unfortunately, only the smaller dataset for image type B was available for this experiment.

The segmentation algorithm alone achieved good results when trained separately for each class (dice score $\approx 95\%$). However, it is clear from visually inspecting the masks (figure 3.11) that further improvements could be made because some artefacts were still visible. These artefacts could be addressed by applying post-processing such as clustering or other traditional computer vision techniques, to identify and remove these small distinct areas.

There were several limitations surrounding the dataset and how it was used. The images were of variable quality and although segmentation was performed in the second approach to remove unwanted artefacts such as vertical streaks and parts of the eyelid, further pre-processing could enhance the dataset. Removing or improving poor quality scans would increase the model accuracies because it would allow the networks to focus only on salient corneal features. Some of the images suffered from blink artefacts where the cornea was barely visible and therefore would not have been good candidates for either segmentation or classification. These images could be completely removed from the dataset by training a classification algorithm to detect good quality corneal images versus noisy images. This algorithm could be applied as an initial step before segmentation. Other images suffered from blurring and these could either be removed from the dataset (identified with a blur metric such as Laplacian variance) or sharpened and included using an algorithm such as blind deconvolution.
It is also worth noting that the overall dataset was relatively small and would therefore benefit from being expanded. This could be achieved by augmentation techniques (e.g. cropping, rotating or scaling) or simply by capturing more images. Future experiments should combine the different image types (both A and B) to maximise the dataset rather using separate models. Type A images capture a narrower section of the cornea compared to type B. An algorithm could be developed to crop and align type B images with type A and this would allow the two datasets to be combined. Alternatively, type B images could simply be resized and fed into the algorithm as a form of data augmentation. Finally, it is also worth noting that within each scan, there were 12 or 25 image slices from the same cornea. This represents repetition and it is likely that if the model correctly classified a slice, then other slices within that scan would also be correctly classified due to them belonging to the same cornea. If this work were to be taken forward to classify a single cornea, it would be logical to exploit the radial nature of the slices by considering all of them simultaneously. This could be done with a simple majority vote classification or with more advanced techniques such as Multi View learning [120]. Although brief experiments were conducted using the latter method, the research could not be completed in time for submission.

Aside from limitations surrounding the datasets, both the classification model common to both approaches and the segmentation model used in Approach 2 could be enhanced by improving the respective network, training for longer or experimenting further with hyperparameters.

In this chapter, a straightforward task was chosen as a proof of concept and as mentioned in the introduction, it does not have clinical utility as a standalone algorithm. However, this work constitutes a first step towards showing how can AI be applied to raw AS-OCT images for keratoconus classification. The most impactful diagnostic algorithm for which raw images are most likely to be of use would be an algorithm to classify subclinical keratoconus from normal eyes, allowing patients to receive care at the nascent stages of disease. Therefore, future research should focus efforts on procuring a large dataset of both normal and clinician labelled subclini-
cal keratoconus eyes to allow this to come to fruition. An easier extension would be to train a model capable of distinguishing between all four classes of manifest keratoconus severity. Keratoconus staging is useful because it allows patients to be triaged more efficiently, reducing the burden on ophthalmic healthcare professionals. It should be noted however, that keratoconus staging can already be achieved by examining the standard outputs of tomographic imaging devices [31].
Chapter 4

Pre CXL Progression
4.1 Introduction

Corneal collagen cross-linking (CXL) by topical application of riboflavin followed by irradiation of the cornea with UV-A light can arrest progression of keratoconus in 88% to 100% of eyes, even when there is relatively advanced disease [35, 36, 37, 38, 39, 40, 121]. The potential benefit of CXL is to minimise the chance of further visual loss with a treatment that has a relatively low risk of complications, and which is cost effective for healthcare providers [57, 58, 59]. However, CXL is usually not offered to all patients at presentation because the disease may have already stabilised. In the recent KERALINK study, 43% of children <17 years of age at presentation had not progressed after 18 months [121]. The definition of progression also varies with the severity of keratoconus, but for early disease a common threshold is either an increase in the maximum keratometry (Kmax) of >1 dioptre, a change in the manifest refractive spherical equivalent of >0.50 dioptre, or an increase in manifest refractive cylinder of >1 dioptre [35]. As mentioned in chapter 1, the ability to predict the trajectory of progression at the first visit would be beneficial for both patients and clinicians. In this chapter, a personalised model to predict keratoconus progression before CXL was generated.

4.2 Methods

A time-to-event model that predicts the probability of an individual progressing to the stage where CXL treatment would be offered is generated. In addition to this, a separate analysis was performed on a subset of the dataset which had genetic data, in the form of single-nucleotide polymorphisms (SNP), with the aim of understanding whether genetic data may be used to augment a keratoconus progression model.

4.2.1 Cohort

From the electronic health record database (OpenEyes) patients aged 13 years and older (mean [SD] age, 28.3 [7.1] years) diagnosed with keratoconus, or suspected
of having keratoconus, attending the Early Keratoconus Clinic (EKC) at Moorfields Eye Hospital between 26th January 2011 and 12th November 2020.

Clinical data included keratometry (Kmax, Front K1, Front K2, Back K1, Back K2), and pachymetry (minimum corneal thickness) captured by Scheimpflug tomography (Pentacam HR; Oculus GmbH). Only scans with a quality score of “good” or “OK,” were included and where multiple scans were taken on the same day, the mean value was used. The date of all CXL procedures was recorded. The protocol for offering CXL throughout the study period was as follows: (1) a documented history prior to referral to the Early Keratoconus Clinic of significant recent disease progression [40], (2) a change in contemporary measurements of 95% above the repeatability limits of the baseline measurements as shown in table 4.1 [40] or (3) a patient considered by a clinician to be at high risk of progression despite their not fulfilling the above two criteria. Exclusion criteria included pregnancy or breastfeeding, uncontrolled ocular surface disease, or a minimum corneal thickness less than 375 µm.

<table>
<thead>
<tr>
<th>Early (Kmax &lt; 55D) (1 or more)</th>
<th>Moderate/advanced (Kmax ≥ 55D) (1 or more)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1D increase Kmax</td>
<td>≥ 2.5D increase Kmax</td>
</tr>
<tr>
<td>≥ 1D increase Front K1 or K2</td>
<td>≥ 2.5D increase Front K1 or K2</td>
</tr>
<tr>
<td>≥ 0.5D increase Back K2</td>
<td>≥ 22 µm decrease Pachymetry</td>
</tr>
<tr>
<td>≥ 16 µm decrease Pachymetry</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4.1:** Defining Disease Progression in Early and Moderate/Advanced Keratoconus Prior to Corneal Cross-Linking. D = Dioptres, Kmax = maximum keratometry, Front K1 = flat keratometry in the central 3mm zone, Front K2 = steep keratometry in the central 3mm zone. Back K2 = steepest posterior keratometry

All the data used for model fitting was taken from the first appointment. Patient demographic data including age, gender, smoking status (current or ex/non-smoker) ethnicity and postcode were available. Ethnicity was coded as 1 for ‘Black’ or ‘Asian’ and 0 for any other category (excluding missing values). Before model fitting,
pachymetry was divided by 10 to generate a meaningful scale. When handling missing data, the main results used complete case analysis (eyes with any missing data were excluded). However, multiple imputation was also explored which avoids data exclusion by generating multiple versions of the dataset with missing values replaced with values sampled from an appropriate distribution. The newly generated versions of the dataset were analysed and the results were then combined and reported.

To explore whether genetic data can help predict keratoconus progression, 28 candidate SNPs were used from a recent keratoconus genome-wide association study [27] containing 926 patients from Moorfields Eye Hospital. The SNP data is encoded as either 0 (homozygous reference genotype), 1 (heterozygous genotype) or 2 (homozygous variant genotype). This is known as additive encoding because 0 refers to no genetic variant, 1 refers to a single variant and 2 refers to a situation where both genes have variants, thus the risk of disease increases additively with the degree of genetic variation [122].

Anonymised data were retrieved and exported to Excel software for analysis (version 15.24 2016, Microsoft Corp.). The study protocol was reviewed and approved by the Clinical Audit Assessment Committee of Moorfields Eye Hospital NHS Foundation Trust (reference CA17/CED/03). Institutional Review Board (IRB) approval was obtained and individual patient consent was not required. The study conformed to the tenets of the Declaration of Helsinki.

4.2.2 Model Fitting and Covariate Selection

A Royston-Parmar flexible parametric survival model was fitted to the data to predict the probability of an eye progressing to CXL [123]. Initial analysis of the covariates was performed by univariate analysis using the same model characteristics as the multivariable model. When selecting covariates for the final multivariable model, backwards stepwise selection was used with a significance level of 0.05. Linear covariates were used for ease of interpretation of the final model. To create a more parsimonious model, the effect on explained variation and discrimination of removing single variables from the model was examined.
4.2.3 Keratometric Progression Sensitivity Analysis

A sensitivity analysis in which keratometric progression was used as an alternative end point was performed. Keratometric progression was defined using thresholds from Gore and associates [40]. When using numerical thresholds to define progression, the appointments for eyes beyond the date of CXL cannot be used. However, censoring these eyes at the date of CXL represents informative censoring. Based on the recommendations of Clarke and associates [124] for investigating the impact of informative censoring, a “best case” data set where eyes were censored at the CXL date and a “worst case” data set where patients were assumed to progress at the CXL date was generated.

The corresponding Kaplan-Meier curves were plotted to provide a visual comparison of the 2 data sets. A Royston-Parmar model was then fitted on both data sets. The same techniques (backward stepwise selection, significance level of 0.05) were used as described in the previous section to fit the model and compare the explained variation and hazard ratios.

4.2.4 Multivariable Model Validation

The model was validated using internal-external cross validation (IECV) [125, 126] in which the dataset was split by geographical region. For the kth region, the model is fitted on the full dataset excluding region k and then Kaplan-Meier curves and predicted survival curves were generated for region k. Seven geographical regions were created based on the patient’s postcode as shown in figure 4.1. To quantitatively assess the validation, Royston and Sauerbrei’s D statistic [127] was calculated for both the model fitted from data excluding region k ($D(k)$) and also the model applied to region k ($D_k$). The difference between these two discrimination metrics ($D_k - D(k)$) was calculated with its corresponding standard error to assess the predictive ability of the model. To demonstrate how the model could be used in practice, three hypothetical patient profiles were chosen with different progression risk profiles (high, medium, low risk) and the predicted time-to-event curves for each were plotted.
Greater London was split into 6 postcode regions around its centre. Any postcodes outside of greater London were classified as the 7th region (“other”). The number of participants per region is given in Table 4.10.

4.2.5 Statistical Analysis

The event of interest was defined as the date that the patient underwent CXL. The time-to-event was calculated as the difference between the first appointment and the date of CXL (or the last patient appointment in the case of censoring). Because there were paired observations (eyes), variance-corrected models to account for correlation between eyes and to ensure that robust standard errors were produced. The choice of scale and selection of degrees of freedom for the Royston-Parmar model was informed by inspecting the Akaike information criterion (AIC) and Bayes information criterion (BIC) [53]. Royston-Parmar models can be fit on a number of different scales (Hazard, Odds, Normal) and the degrees of freedom for the baseline spline can also take a range of integer values. When selecting these two aspects of the model, it is important not to assume linear covariate effects for continuous variables. Therefore, each continuous variable was categorised into 5 groups first.
<table>
<thead>
<tr>
<th>DF</th>
<th>Hazard AIC</th>
<th>Hazard BIC</th>
<th>Odds AIC</th>
<th>Odds BIC</th>
<th>Normal AIC</th>
<th>Normal BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>953.51</td>
<td>1113.8</td>
<td>724.22</td>
<td>884.51</td>
<td>683.78</td>
<td>844.06</td>
</tr>
<tr>
<td>2</td>
<td>483.05</td>
<td>649.06</td>
<td>416.68</td>
<td>582.69</td>
<td>511.2</td>
<td>677.21</td>
</tr>
<tr>
<td>3</td>
<td>330.09</td>
<td>501.82</td>
<td>307.59</td>
<td>479.32</td>
<td>512.5</td>
<td>684.24</td>
</tr>
<tr>
<td>4</td>
<td>305.77</td>
<td>483.23</td>
<td>234.19</td>
<td>411.65</td>
<td>283.45</td>
<td>460.91</td>
</tr>
<tr>
<td>5</td>
<td>67.06</td>
<td>250.25</td>
<td>25.8</td>
<td>208.98</td>
<td>88.06</td>
<td>271.24</td>
</tr>
<tr>
<td>6</td>
<td>108.82</td>
<td>297.72</td>
<td>62.16</td>
<td>251.07</td>
<td>119.89</td>
<td>308.8</td>
</tr>
</tbody>
</table>

Table 4.2: Table used to choose the scale (Hazard, Odds, Normal) and degrees of freedom for the baseline spline function (DF) for the Royston Parmar model. Each continuous variable was categorized into 5 groups and then every combination of model (3 scales x 6 degrees of freedom = 18 models) was fitted and evaluated in terms of AIC and BIC. AIC = Akaike information criterion, BIC = Bayes information criterion.

The investigation involved iterating over the 3 different scales and 6 degrees of freedom to find the optimal (minimum) AIC and BIC. The results of iterating over both scale and degrees of freedom to guide further analysis can be seen in table 4.2. 5 degrees of freedom was chosen because it gave the lowest AIC and BIC for all scales. Although the Odds scale gave the lowest AIC and BIC, it was decided to use the Hazards scale because of its familiarity. Royston and Sauerbrei’s D statistic was used as a measure of discrimination and $R^2D$ as a measure of explained variation (both calculated on the natural scale of the model). All of the primary results were generated from a complete case analysis, however, an additional analysis was performed using multiple chained imputation (predictive mean matching approach with 5 nearest neighbours). Model fitting was performed in Stata 13 (StataCorp LP, Texas, USA) and the Royston-Parmar model was fitted using the stpm2 package.
4.3 Results

4.3.1 Cohort

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Type</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
<th>Missing No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front K1 (D)</td>
<td>Numeric</td>
<td>45.31</td>
<td>3.86</td>
<td>8,813</td>
<td>528 (5.7)</td>
</tr>
<tr>
<td>Front K2 (D)</td>
<td>Numeric</td>
<td>48.39</td>
<td>4.85</td>
<td>8,839</td>
<td>502 (5.4)</td>
</tr>
<tr>
<td>Back K1 (D)</td>
<td>Numeric</td>
<td>-6.53</td>
<td>0.75</td>
<td>7,949</td>
<td>1392 (14.9)</td>
</tr>
<tr>
<td>Back K2 (D)</td>
<td>Numeric</td>
<td>-7.23</td>
<td>0.93</td>
<td>8,702</td>
<td>639 (6.8)</td>
</tr>
<tr>
<td>Kmax (D)</td>
<td>Numeric</td>
<td>54.14</td>
<td>8.01</td>
<td>8,834</td>
<td>507 (5.4)</td>
</tr>
<tr>
<td>Pachymetry (µm)</td>
<td>Numeric</td>
<td>462.92</td>
<td>46.15</td>
<td>8,946</td>
<td>395 (4.2)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>Numeric</td>
<td>28.28</td>
<td>7.10</td>
<td>9341</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Genetic data(^a)</td>
<td>Ordinal</td>
<td>N/A</td>
<td>N/A</td>
<td>1141</td>
<td>8020 (85.9)</td>
</tr>
<tr>
<td>Ethnicity(^b)</td>
<td>Categorical</td>
<td>N/A</td>
<td>N/A</td>
<td>4889</td>
<td>4452 (47.7)</td>
</tr>
<tr>
<td>(40.1% white)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male gender</td>
<td>Categorical</td>
<td>N/A</td>
<td>N/A</td>
<td>9341</td>
<td>0 (0)</td>
</tr>
<tr>
<td>(67% male)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoker(^c)</td>
<td>Categorical</td>
<td>N/A</td>
<td>N/A</td>
<td>9341</td>
<td>0 (0)</td>
</tr>
<tr>
<td>(4.5% smoker)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: Summary Statistics for the Available Covariates at the First Examination for 9341 Eyes Recorded at First Visit. Back K1 = flattest posterior keratometry, Back K2 = steepest posterior keratometry, Front K1 = flattest anterior keratometry, Front K2 = steepest anterior keratometry, Kmax = maximum keratometry, pachymetry = minimum corneal thickness, N/A = not applicable.

\(^a\)Genetic data composed of 28 SNPs and was encoded in an additive fashion (0, 1, 2).
\(^b\)1 = Black or Asian, 0 = otherwise.
\(^c\)0 = nonsmoker/ex-smoker, 1 = current smoker.
From a potential of 9,341 eyes (4316 pairs of eyes and 709 individual eyes), the final model used 8,701 eyes of 4,823 patients, with 3,232 eyes that had CXL. The mean age was 28.3 years with standard deviation of 7.1 years. 640 eyes were excluded with missing data however multiple imputation results are included at the end of this section. Table 4.3 summarises the available covariates along with missing data percentages.

### 4.3.2 Model Fitting and Covariate Selection (Genetic Data)

Eyes with genetic data were analysed separately because this data was only available for 14% of patients. Of 926 patients (1852 eyes) with genetic data, 531 eyes were excluded with incomplete keratometry or CXL data, which left 1321 eyes, of which 665 had CXL. With univariate analysis of the 28 SNPs only rs72631889 was found to be significant (P=0.01) (table 4.4). A multivariable model was then produced via backwards selection on this subset of eyes using corneal data, patient data and rs72631889 as an additional covariate as shown in table 4.5. However rs72631889, although significant (P=0.005), had a negligible contribution (0.3%) to the explained variation in the final model.
<table>
<thead>
<tr>
<th>SNP</th>
<th>Odds Ratio [95% CI]</th>
<th>$R^2_D$</th>
<th>$D$</th>
<th>$R^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs116390263</td>
<td>0.96 [0.62; 1.49]</td>
<td>0.85</td>
<td>0</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>rs116452738</td>
<td>0.66 [0.24; 1.86]</td>
<td>0.43</td>
<td>0.01</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>rs3131972</td>
<td>1.06 [0.93; 1.21]</td>
<td>0.35</td>
<td>0</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>rs3131962</td>
<td>1.08 [0.94; 1.24]</td>
<td>0.29</td>
<td>0</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>rs13303222</td>
<td>1.02 [0.86; 1.2]</td>
<td>0.85</td>
<td>0</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>rs4970382</td>
<td>1.04 [0.93; 1.18]</td>
<td>0.48</td>
<td>0</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>rs116587930</td>
<td>1.26 [0.91; 1.75]</td>
<td>0.17</td>
<td>0.01</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>rs11240779</td>
<td>1.11 [0.98; 1.26]</td>
<td>0.1</td>
<td>0</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>rs4970459</td>
<td>0.98 [0.84; 1.14]</td>
<td>0.79</td>
<td>0</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>rs12562034</td>
<td>1 [0.84; 1.2]</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>rs116720794</td>
<td>1.09 [0.84; 1.42]</td>
<td>0.52</td>
<td>0</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>rs13303101</td>
<td>1.1 [0.94; 1.29]</td>
<td>0.24</td>
<td>0</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>rs4970383</td>
<td>1.06 [0.93; 1.2]</td>
<td>0.37</td>
<td>0</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>rs79373928</td>
<td>1.45 [0.82; 2.58]</td>
<td>0.21</td>
<td>0.01</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>rs192998324</td>
<td>1.15 [0.72; 1.83]</td>
<td>0.57</td>
<td>0</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>rs57181708</td>
<td>1.05 [0.88; 1.26]</td>
<td>0.55</td>
<td>0</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4.4:** Univariable results for genetic data. This includes the 28 most significantly associated Single Nucleotide Polymorphisms (SNPs). $D =$ Royston and Sauerbrei’s $D$ statistic (used as a measure of discrimination), $R^2_D =$ explained variation, SNP = single-nucleotide polymorphism. This table includes the 28 most significantly associated SNPs.
<table>
<thead>
<tr>
<th>Covariate</th>
<th>Hazard Ratio [95% CI]</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at presentation</td>
<td>0.92 [0.90; 0.93]</td>
<td>0.001</td>
</tr>
<tr>
<td>Kmax</td>
<td>1.10 [1.07; 1.12]</td>
<td>0.001</td>
</tr>
<tr>
<td>Back K2</td>
<td>1.60 [1.36; 1.87]</td>
<td>0.001</td>
</tr>
<tr>
<td>Pachymetry(^a)</td>
<td>0.96 [0.94; 0.99]</td>
<td>0.010</td>
</tr>
<tr>
<td>rs72631889</td>
<td>0.42 [0.23; 0.77]</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 4.5: Hazard ratios for genetic sub-analysis model using 1144 eyes containing clinical data and rs72631889 SNP. Back K2 = steepest posterior keratometry, Kmax = maximum Keratometry.

\(^a\)Minimum pachymetry in steps of 10\(\mu\)m.

4.3.3 Model Fitting and Covariate Selection (Excluding Genetic Data)

The results of the univariate time-to-event analysis on the hazards scale using a Royston-Parmar flexible parametric model is shown in table 4.6. Genetic data was excluded from this analysis. All variables except smoking status were significant.

The explained variation \((R^2 D)\) and discrimination (D) were highest for age (17\%) and Kmax (15\%) with Front K1, Front K2, Back K1, Back K2 and pachymetry each explaining 6-10\% of the variation. Notably, gender and ethnicity, although significant in the univariate analysis, did not contribute to explained variation. The hazard ratios for significant covariates indicate that increasing age at presentation, greater pachymetry and flatter (less negative) posterior keratometry values decrease risk of having CXL, whilst steeper anterior keratometry values and male gender increase the risk of having CXL.
### Table 4.6

Table 4.6: Univariable and final multivariable model for all considered covariables excluding genetic data in the training dataset fitted on the hazards scale with 5 degrees of freedom. $R^2D$ = explained variation, $D$ = Royston and Sauerbrei’s D statistic (used as a measure of discrimination). Kmax = maximum keratometry, Front K1 = flattest anterior keratometry, Front K2 = steepest anterior keratometry, Back K1 = flattest posterior keratometry, Back K2 = steepest posterior keratometry, Pachymetry = minimum corneal thickness.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Hazard Ratio [95% CI]</th>
<th>P-Value</th>
<th>R$^2$D</th>
<th>D</th>
<th>Hazard Ratio [95% CI]</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnicity</td>
<td>1.14 [1.02; 1.27]</td>
<td>0.02</td>
<td>0.4%</td>
<td>0.13</td>
<td>0.9 [0.90; 0.91]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Smoker$^1$</td>
<td>1.07 [0.9; 1.28]</td>
<td>0.46</td>
<td>0.1%</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male Gender</td>
<td>1.11 [1.01; 1.21]</td>
<td>0.02</td>
<td>0.2%</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at presentation</td>
<td>0.91 [0.9; 0.92]</td>
<td>&lt; 0.001</td>
<td>16.7%</td>
<td>0.92</td>
<td>0.9 [0.90; 0.91]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Kmax</td>
<td>1.06 [1.05; 1.06]</td>
<td>&lt; 0.001</td>
<td>14.9%</td>
<td>0.86</td>
<td>1.08 [1.07; 1.09]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Front K1</td>
<td>1.09 [1.08; 1.1]</td>
<td>&lt; 0.001</td>
<td>7.0%</td>
<td>0.56</td>
<td>0.93 [0.91; 0.94]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Front K2</td>
<td>1.08 [1.07; 1.08]</td>
<td>&lt; 0.001</td>
<td>9.8%</td>
<td>0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back K1$^2$</td>
<td>0.67 [0.64; 0.71]</td>
<td>&lt; 0.001</td>
<td>5.9%</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back K2$^2$</td>
<td>0.7 [0.67; 0.72]</td>
<td>&lt; 0.001</td>
<td>8.4%</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pachymetry$^3$</td>
<td>0.93 [0.92; 0.94]</td>
<td>&lt; 0.001</td>
<td>7.5%</td>
<td>0.58</td>
<td>0.95 [0.93; 0.96]</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

$^1$nonsmoker/ex-smoker; 1 = current smoker.
$^2$Back K1 and Back K2 are negative values such that patients with advanced keratoconus are typically associated with large negative values.
$^3$Minimum pachymetry in steps of 10 µm.
After fitting a multivariable model, the significant covariates were age, Kmax, Front K1, Front K2 and pachymetry. When single variables were individually removed from the model whilst keeping the rest, the effect this had on explained variation and discrimination is shown in table 4.7. Removing age has the most dramatic effect, reducing $R^2D$ from 33.3% to 19.7%. Removing Kmax reduced $R^2D$ from 33.3% to 28.4%. The other covariates are less important with pachymetry, Front K1 and Front K2 reducing $R^2D$ by <1% when removed. A model without Front K2 was chosen on the basis of parsimony, which was supported by the fact that Front K1 and Front K2 were highly correlated (R2=0.91) as shown in figure 4.2. The final fitted model hazard ratios can be seen on the multivariable column of table 4.6. It is notable that an increase in Front K1 now has a protective effect in the final model. The explained variation and discrimination for the final model were 32.7% and 1.43 respectively [128]. The opposing effect of Kmax and Front K1 can be explained by examining their regression coefficients before converting to hazard ratios; Kmax has a positive coefficient (0.0795) and Front K1 has a negative coefficient (-0.0749). This is logically similar to including the combined covariate (Kmax - Front K1) in the model which can be viewed clinically as a proxy for irregular astigmatism. Combining Front K1 and Front K2 into a single covariate as (Front K2-Front K1) (standard definition of astigmatism) was also investigated, but the corresponding p-value was not significant.
Table 4.7: The Effect on Explained Variation and Discrimination of Removing Individual Covariates From the Full Model. $R^2D = \text{explained variation, } D = \text{Royston and Sauerbrei’s } D \text{ statistic (used as a measure of discrimination), } \text{Front K1 = flattest anterior keratometry, } \text{Front K2 = steepest anterior keratometry, } K_{\text{max}} = \text{maximum Keratometry.}$

<table>
<thead>
<tr>
<th>Covariate Removed</th>
<th>$R^2D$</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.333</td>
<td>1.45</td>
</tr>
<tr>
<td>Age at presentation</td>
<td>0.197</td>
<td>1.01</td>
</tr>
<tr>
<td>K_{\text{max}}</td>
<td>0.284</td>
<td>1.29</td>
</tr>
<tr>
<td>Front K1</td>
<td>0.324</td>
<td>1.42</td>
</tr>
<tr>
<td>Front K2</td>
<td>0.327</td>
<td>1.43</td>
</tr>
<tr>
<td>Pachymetry$^a$</td>
<td>0.324</td>
<td>1.42</td>
</tr>
</tbody>
</table>

$^a$Minimum pachymetry in steps of 10µm

Figure 4.3 visually depicts the result of applying the final model to the original dataset. As expected, the predicted mean survival curves closely follow the Kaplan-Meier curves. To demonstrate the use of the model in clinical practice, survival curves for three hypothetical patients followed for five years are shown in figure 4.4. A web application has also been produced from the model which can be accessed at http://beta.moorfieldsxcl.com.
**Figure 4.2:** Correlations among anterior keratometry covariates and pachymetry. Front K1 and Front K2 can be seen to be very highly correlated. This correlation is exemplified by the fact that removing one or other of these covariates has a negligible effect on the explained variation of the model. Given this high correlation, Front K2 was excluded from the final multivariable model. \( k_{\text{max}} = \) maximum keratometry, \( \text{front}_k1 = \) flattest anterior keratometry, \( \text{front}_k2 = \) steepest anterior keratometry, \( \text{pachy} = \) thinnest point pachymetry.
Figure 4.3: Chart showing how the Royston-Parmar model fits the entire dataset. The eyes were split into 4 risk groups by their prognostic index: <25th centile (low risk), 25-50th centile (medium-low risk), 50-75th centile (medium-high risk), >75th centile (high risk). The number of eyes at risk corresponds to the Kaplan-Meier curves.
**Figure 4.4:** Time-to-event curves that predict the risk of progression to CXL for three hypothetical patient profiles. The blue line represents a high risk patient who has a 95% probability of progressing to CXL at 5 years. The red line is a medium risk patient who has a 48% probability of progressing to CXL at 5 years. The green line is a low risk patient who has a 14% probability of progressing to CXL at 5 years. Abbreviations: pachy, pachymetry.
4.3.4 Keratometric Progression Sensitivity Analysis

By examining the Kaplan-Meier curves in 4.5, it can be seen that the best case time-to-event curve indicates a 40% survival probability at 5 years whilst the worst case curve indicates a 27% survival probability at 5 years. This 13% difference in survival probability at 5 years represents the upper bound of the discrepancy in survival probability within the data. After fitting the Royston-Parmar model, amongst the hazard ratios which overlap (age, Kmax, Front K2), there was reasonable similarity (tables 4.8 and 4.9). Most importantly, the model fitted to the best case had an explained variation of 11% compared to 23% for the worst case indicating a significant difference in model performance depending on the assumptions used for handling eyes which received CXL.

![Best vs Worst case survival](image)

**Figure 4.5:** Kaplan-Meier time-to-event curves for keratometric progression in the best (censored at CXL date) and worst (progressed at CXL date) case scenarios.
### Table 4.8: Best Case Hazard Ratios. Kmax = maximum keratometry, Front K2 = steepest anterior keratometry.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Hazard Ratio</th>
<th>P-Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.95</td>
<td>0.001</td>
<td>[0.95; 0.96]</td>
</tr>
<tr>
<td>Kmax</td>
<td>1.08</td>
<td>0.001</td>
<td>[1.06; 1.10]</td>
</tr>
<tr>
<td>Front K2</td>
<td>0.91</td>
<td>0.001</td>
<td>[0.88; 0.93]</td>
</tr>
</tbody>
</table>

### Table 4.9: Worst Case Hazard Ratios. Kmax = maximum keratometry, Front K1 = flattest anterior keratometry, Front K2 = steepest anterior keratometry.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Hazard Ratio</th>
<th>P-Value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.93</td>
<td>0.001</td>
<td>[0.92; 0.94]</td>
</tr>
<tr>
<td>Kmax</td>
<td>1.09</td>
<td>0.001</td>
<td>[1.08; 1.10]</td>
</tr>
<tr>
<td>Front K1</td>
<td>0.98</td>
<td>0.02</td>
<td>[0.96; 1.00]</td>
</tr>
<tr>
<td>Front K2</td>
<td>0.95</td>
<td>0.002</td>
<td>[0.93; 0.97]</td>
</tr>
<tr>
<td>Pachymetry</td>
<td>0.97</td>
<td>0.003</td>
<td>[0.96; 0.98]</td>
</tr>
</tbody>
</table>

*Minimum pachymetry in steps of 10 µm.

#### 4.3.5 Multivariable Model Validation

When performing validation using internal-external cross validation, figure 4.6 shows the ability of the final model to predict keratoconus progression across different geographic regions. No significant differences in prognostic factors across regions were identified. The model prediction curves generally follow the Kaplan-Meier curves. Notably, region 5 (South West Greater London) and region 7 (other regions) have a worse predictive performance than the other regions, indicating that these regions have different characteristics compared with the remainder of the dataset used for model fitting. This could be due to differing patient characteristics, such as complex cases that required referral to a tertiary referral centre rather than being managed locally. Overall, the prediction becomes less accurate over time, which is expected due to low numbers with follow-up beyond three years. Table 4.10
displays quantitative validation results of the model using internal external validation. The difference column $D_k - D(k)$ is a measure of predictive ability. Region 7 (other regions outside of Greater London) has the greatest discrepancy in discrimination (-0.26) which indicates that the model fitted when excluding region 7 had greater discriminative ability than when applied to region 7 alone.

Figure 4.6: Predicted and observed survival curves for seven postal code regions of Greater London as shown in Figure 4.1 using IECV. The eyes were split into 4 risk groups by their prognostic index: <25th centile (low risk), 25-50th centile (medium-low risk), 50-75th centile (medium-high risk), >75th centile (high risk).
Table 4.10: Internal-external cross-validation across different geographic regions. $D_k$ = Royston and Sauerbrei’s D statistic for the model applied to region k. $D(k)$ = Royston and Sauerbrei’s D statistic for the model fitted from data excluding region k. D S.E. = Standard error of $D_k - D(k)$.

### 4.3.6 Multiple Imputation Results

When applying multiple chained imputation, all 9341 eyes were included in the analysis and the missing values were imputed for Front K1, Front K2, Back K1, Back K2, Kmax, pachymetry and ethnicity. Imputation was not attempted for genetic data because the percentage missing (86%) was deemed to be too high. When the model fitting process was repeated using multiple chained imputation, the hazard ratios were very similar to the complete case analysis (table 4.11).

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Hazard Ratio (95% CI)</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at presentation</td>
<td>0.91 (0.90, 0.92)</td>
<td>.001</td>
</tr>
<tr>
<td>Kmax</td>
<td>1.10 (1.09, 1.11)</td>
<td>.001</td>
</tr>
<tr>
<td>Pachymetry&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.97 (0.96, 0.98)</td>
<td>.001</td>
</tr>
<tr>
<td>Front K1</td>
<td>0.95 (0.93, 0.98)</td>
<td>.001</td>
</tr>
<tr>
<td>Front K2</td>
<td>0.94 (0.91, 0.96)</td>
<td>.001</td>
</tr>
</tbody>
</table>

Table 4.11: Multiple Chained Imputation Results. Kmax = maximum keratometry, Front K1 = flattest anterior keratometry, Front K2 = steepest anterior keratometry.

<sup>a</sup>Minimum pachymetry in steps of 10µm.
4.4 Discussion

This chapter has incorporated demographic, keratometric, and genetic data to generate a prognostic model of keratoconus progression to CXL. It has been shown that parameters recorded at the first examination (age, Kmax, Front K1, minimum pachymetry) can produce a time-to-event curve to calculate a personalised risk for keratoconus progression. Although time to CXL was chosen rather than keratometric progression as the end point for the time-to-event analysis, a sensitivity analysis using keratometric progression was performed, and it was found that a CXL model accounts for a much higher proportion of the explained variation (33%) compared to the keratometric model (11% or 23% for best and worst case respectively). The opposing effects of Kmax and Front K1 were unexpected, but similar to including the combined covariate (Kmax - Front K1) in the model; a possible explanation is that the opposing effect is the result of an increase in irregular astigmatism. Of the significant covariates in the model, younger age made the greatest contribution. Thus, one should have a lower threshold for treatment in younger patients.

When applying internal-external cross validation, the survival curves closely followed the Kaplan-Meier survival curves for each of the geographic regions, which indicates generalisability, and model discrimination between training and cross validation groups was similar, indicating that the predictive ability is well maintained. Finally, SNP genetic data had limited additional predictive utility for keratoconus progression. However, the genetic dataset was relatively small (926 patients), and recruitment was based on the presence of keratoconus, as opposed to the severity of keratoconus, or any other index of risk of rapid progression.

The Royston-Parmar model has previously been used to predict the likelihood of the worst eye of patients with keratoconus progressing to corneal transplantation [129]. In their final model, they chose 3 significant covariates: Kmax, age and ethnicity. The reported covariate hazard ratios which overlap with this study (Kmax and age) were different in magnitude but in the same direction. When performing internal validation, their model exhibited good predictive ability. They produced time-dependent receiver operating curves using the validation set and found 1 year
sensitivity and specificity to be 92.8% and 94.6% respectively. Using logistic regression, Kato et al. found that the two strongest factors associated with the requirement for CXL were age and Kmax [130], which is consistent with the findings in this chapter.

An ability to generate personalised time-to-event curves that predict progression to CXL (figure 4.4) could directly inform clinical decisions that benefit patient care. Firstly, patients may better understand their own risk for progression and feel more confident in choosing their management options. Secondly, for both clinicians and patients, the prediction of progression may contribute to scheduling treatments, including prioritising patients at high risk of early progression. For example, patients at high risk with a 98% probability of progressing to CXL at 5 years could be offered CXL at the point of first diagnosis without waiting to demonstrate keratometric progression. Medium risk patients may benefit from a period of clinician-led topographic monitoring. For the lowest risk patients, optometry-led monitoring in the community may be sufficient. This risk stratification could be tailored to regions and reflect local needs and resources such as provision of monitoring services in regions with lower risk and greater capacity for CXL in areas with more high risk patients. Finally, when a decision is made to postpone CXL for further monitoring, the time-to-event curve can contribute to decisions on the scheduling of future follow up reviews, with perhaps shorter time periods where the curve is steepest. Recommendations based on this model on clinical practice is yet to be evaluated.

This study is subject to several limitations inherent to the dataset. First, if patients had CXL at another hospital, this may not be reliably recorded in the source database. This could lead to a very small number of patients being included in the analysis who have already had CXL. Second, ethnicity is a well established risk factor for keratoconus and keratoconus progression [131, 132, 133], but ethnicity is now an optional field at patient registration at the EKC and this information was unavailable for approximately 50% of the dataset. However, even when the dataset restricted was to those eyes with ethnicity records, it was not found to be a significant covariate. Third, though the cohort used for univariable and multivariable
analysis were identical, the number of eyes where all covariates were available was lower than for univariable analysis due to missing data. Finally, when multiple imputation was used to generate a multivariable model, ethnicity was still not found to be significant. In the model fitting process, simple backwards selection was used as opposed to the multivariate fractional polynomial (MFP) method [128]. After initial investigations, the results of MFP yielded nonlinear functional forms of the covariates and, whilst this method may have slightly increased the predictive power of the prognostic model, the resulting hazard ratios would be very hard to interpret. In addition, time dependent effects were not examined for the covariates, which may provide a more accurate model fit, and future studies should examine this option. Finally, although no external validation dataset was available, internal external cross validation allowed us to confirm that the model was generalisable across geographical regions.

In conclusion, a prognostic model was fitted for progression of keratoconus to CXL which generates a time-to-event curve using age, Kmax, Front K1, minimum pachymetry from time of presentation. Incorporation of a relatively small genetic dataset does not improve the explained variation of the model. Personalised modelling of risk may improve patients’ understanding of their condition and the need for CXL. Such a model may help better improve patients and aid clinician decision making to CXL to achieve better outcomes and judicious use of healthcare resources.
Chapter 5

Post CXL Progression
5.1 Introduction

In this chapter, the problem of defining progression after the CXL operation is considered. Estimates of CXL treatment success are determined by the definition of progression that is adopted [134, 135, 136, 137], with no consensus on the parameters that should be used. The heterogeneity of definitions used to define progression, as well as inconsistencies or absent reporting of the key outcome measure of arrested progression has been highlighted in a recent Cochrane meta-analysis [138].

The predominant method to identify keratoconus progression is to measure the change in a single parameter at a fixed threshold, such as a change in front K2 greater than 1.5D or a change in Kmax greater than 1D, with the threshold applied across the full spectrum of disease severity. However, as the severity of keratoconus increases, the repeatability of keratometry reduces [4, 5, 6, 7]. A modification that recognises the reduced repeatability associated with higher keratometry values is to incorporate a second larger threshold, typically for keratometry values above 55D. However, with a fixed definition, progression in early keratoconus can be missed if small (<1D), but repeatable changes are ignored. A missed diagnosis of progression can lead to visual loss if CXL is delayed. In addition, greater variability of measurements may result in misclassification as progression in more advanced disease.

This chapter aims to develop adaptive thresholds for progression adapted to the severity continuum of keratoconus. The utility of this method is demonstrated in a large dataset of eyes that were followed for up to six years after CXL (five years beyond a reference visit at 9-15 months after CXL). The results from this method were then compared with results from the same dataset when standard static definitions of progression were used.
5.2 Methods

The protocol was approved as a clinical audit by Moorfields Eye Hospital NHS Foundation Trust (references CA17/CED/03 and 22/PR/0249). It complied with the Interventional Procedure Guidance (IPG466) of the National Institute for Health and Care Excellence (NICE) and adhered to the tenets of the Declaration of Helsinki. In this section, adaptive thresholds for Kmax, Front K2, Back K2 and pachymetry are defined using a variation of Bland Altman analysis. These adaptive thresholds are then used in three new definitions of progression and compared to three existing definitions of progression using Kaplan-Meier analysis.

Minimum pachymetry was excluded as a parameter for progression due to the uncertainty that arises from stromal compaction and remodelling in the first months after CXL, the thinner and more homogeneous epithelial layer, and the reduced reliability of Scheimpflug measurements due to light scatter [139, 140, 141, 142]. However pachymetry was included in the evaluation of the 95% LoA since it may be clinically relevant when assessing suitability for repeat CXL.

5.3 Data

The dataset consisted of 17396 eyes from 8807 keratoconus patients which visited the EKC between the 26th January 2011 and the 2nd of September 2021. This dataset consisted of eyes before and after the CXL operation. Demographic data included age, gender, and self-identified ethnicity. Slit-lamp examination and Scheimpflug corneal tomography were performed at 6 and 12 months after treatment and then annually, with data added to an electronic spreadsheet. However, depending on the stage of analysis, different filtering was applied and this is explained in the respective sections below.
5.3.1 Adaptive Thresholds

To define the keratometry adaptive thresholds, the entire dataset (consisting of eyes before and after CXL) was used. Only eyes with repeated scans at the same appointment were considered. The first two suitable scans were selected from each appointment with an adequate quality score. This resulted in 6463 eyes with paired readings of Kmax, Front K2, Back K2, and pachymetry with which to define the thresholds. The difference and the average between all the pairs of readings were calculated. A variation of Bland Altman analysis was used to generate regression-based limits of agreement (LoA) for the whole range of keratoconus severity for three keratometry values and pachymetry. This variation is an accepted method when the standard deviation of repeated measurements varies according to the magnitude of the measurement [143, 144]. The steps to achieve this are as follows:

- Regress the differences on averages:

\[
\text{Difference} = a + b(\text{Average})
\]  
(5.1)

- Calculate the absolute residuals between the observed and predicted differences:

\[
\text{AbsResidual} = |\text{Difference}_{\text{obs}} - \text{Difference}_{\text{pred}}|
\]

- Regress the absolute residuals on the averages:

\[
\text{AbsResidual} = a + b(\text{Average})
\]  
(5.2)

- Multiply coefficients from equation 5.2 by \(\sqrt{\frac{\pi}{2}}\) to get an equation for the standard deviation:

\[
\text{StdDev} = a\sqrt{\frac{\pi}{2}} + b\sqrt{\frac{\pi}{2}}\text{Average}
\]

- The 95% LoA are therefore given as:

\[
\text{LoA}_{95} = \pm 1.96\text{StdDev}
\]
5.3.2 Post CXL Progression Definition

Three new definitions of progression post CXL were created from the adaptive thresholds:

- Increase in K2 > $LoA_{95}$
- Increase in Kmax > $LoA_{95}$
- An increase in at least 2 parameters beyond the $LoA_{95}$ from: Kmax, Front K2, Back K2.

The increase was measured against a baseline appointment defined as the first appointment 9-15 months after the CXL operation. An eye was said to have progressed at the first appointment T after the baseline appointment where the increase between the baseline appointment and appointment T was greater than $LoA_{95}$.

Kaplan-Meier survival curves were plotted for each definition. These three survival curves based on the adaptive definitions were compared to three corresponding curves based on the commonly used fixed (non-adaptive) thresholds of $\geq 1$D increase in Kmax, $\geq 1.5$D increase in K2, and a dual threshold multi-parameter definition in which eyes were also grouped according to their keratometry as above or below 55D [135]. The dual threshold multi-parameter definition is an amalgamation of the progression definition used in an earlier report, but adapted with more recently published repeatability limits [135, 145]. All six definitions are summarised in Table 5.1.

Eyes were included if the participants were $\geq 13$ years of age and had epithelium-off accelerated pulsed high-fluence CXL [30 mW/cm², 1.5 sec on/off cycle, energy 7.2 J/cm² (Avedro KXL, San Clemente, USA)] performed between January 2014 and September 2021. Both eyes were included if they both had CXL. Exclusion criteria for CXL included pregnancy or breastfeeding, uncontrolled ocular surface disease, or a minimum corneal thickness measurement of less than 375 µm. Patients with pellucid marginal corneal degeneration and other ectatic corneal disorders were also excluded. After these initial exclusions, there were 5035 eyes remaining.
In order to generate the survival datasets for Kaplan-Meier analysis, all appointments without \text{Kmax}, \text{Front K2} and \text{Back K2} were removed from the dataset. Eyes that did not have at least one valid appointment during and after the 9-15 month baseline period were excluded from the analysis. After these subsequent exclusions, 1677 eyes from 1217 patients were available for analysis. The reasons for exclusion are summarised in Figure 5.1.

![Flow chart for exclusion process used when generating survival dataset](image)

**Figure 5.1**: Flow chart for exclusion process used when generating survival dataset

The filtered dataset was then transformed into definition-specific survival datasets by marking an eye as experiencing the event when it crossed the corresponding threshold. If it never crossed the threshold, it was either censored at the last available appointment date, or at five years after the baseline appointment. Each definition resulted in a different survival dataset.

### 5.3.3 Statistical Methods

Bland Altman analysis was performed to generate regression based 95\% LoA in RStudio version 1.3 (https://www.rstudio.com). Non-parametric estimation of the risk of keratometric progression was performed using the Kaplan-Meier method (Stata 13, StataCorp LP, Texas, USA). McNemar’s test was used to compare corresponding pairs of eyes at five years following the reference visit 9-15 months after CXL to establish whether the results from any two definitions of progression were statistically significant.
<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-adaptive</td>
<td>Single threshold Kmax [136, 146]</td>
<td>Increase in Kmax $\geq$ 1D</td>
</tr>
<tr>
<td></td>
<td>Single threshold K2 [134]</td>
<td>Increase in K2 $\geq$ 1.5D</td>
</tr>
</tbody>
</table>
|                     | Dual threshold multi parameter [135, 145] | If Kmax $<55$D an increase in 2 parameters from:  
Kmax $\geq$ 1D, Front K2 $\geq$ 1D, Back K2 $\leq$ -0.25D  
If Kmax $\geq$ 55D an increase in 2 parameters from:  
Kmax $\geq$ 2.5D, Front K2 $\geq$ 2.5D, Back K2 $\leq$ -0.5D |
| Adaptive            | Single threshold K2           | Increase in K2 $>95\%$ repeatability limit                                   |
|                     | Single threshold Kmax         | Increase in Kmax $>95\%$ repeatability limit                                 |
|                     | Multi parameter threshold     | An increase in 2 parameters beyond the 95% repeatability limit from:  
Kmax, Front K2, Back K2                                                  |

Table 5.1: Definitions of keratoconus progression based on fixed or adaptive thresholds.
significantly different.

5.4 Results

5.4.1 Adaptive Thresholds

**Figure 5.2:** The v-shaped Bland Altman charts with 95% LoA (red dashed lines) for paired readings of Kmax, front K2, back K2, and pachymetry. The blue dashed lines on the front K2 and Kmax plots represent the single thresholds as defined in the main text. The rug plots on the vertical and horizontal axes of each chart show the frequency of the data. For Kmax and front K2, the v-shaped LoA increases as the mean keratometry value in dioptres increases. For back K2, the v-shaped LoA increases as the mean keratometry decreases. For Kmax and front K2 the v-shaped LoA appear to be a better fit for the data compared to a single threshold. The LoA for thinnest pachymetry is also greater for thinner corneas.

The p-value associated with the first regression (equation 5.1) to generate the 95%
LoA was not statistically significant for all corneal parameter (P>0.05). The mean of the differences was therefore used to estimate $Difference_{pred}$, which was close to zero in all cases [143, 147]. The p-value for the second regression (equation 5.2) was highly significant for all corneal parameters (P<0.001). The resulting Bland Altman charts for Kmax, Front K2, Back K2 and pachymetry are shown in Figure 5.2. The red dashed lines represent the 95% LoA, which are v-shaped and increasing as the absolute value of mean Kmax, front K2 and back K2 increases, suggesting that the repeatability of the readings is reduced.
5.4.2 Post CXL Progression Definition

Figure 5.3: Kaplan-Meier survival curves based on a dataset of 1677 eyes using 6 different definitions of progression. The baseline examination was the first appointment between nine and fifteen months after corneal cross-linking (CXL). The five-year progression rate was 6% for the dual threshold multi-parameter method, 8% for the single $\geq 1.5D$ threshold K2, 20% for the adaptive threshold multi parameter method, 22% for the adaptive threshold K2 method, 22% for the adaptive threshold Kmax, and 29% for the single $\geq 1D$ threshold Kmax method. Confidence intervals for these survival curves are shown in Figure 5.4.

Of the 1217 included patients, 72% were male, and the self-declared ethnicity was non-white in 373 (40%) of the 934 patients with available data. The mean value for Kmax in the included eyes was 57.7D (standard deviation (SD) 8.0D, range 42.8D to 92.4D), 50.2D for front K2 (SD 5.2D, 36.7D to 74.4D), -7.6D for back K2 (SD 1.0D, -12.0D to 5.3D), and 447.8 $\mu$m for minimum pachymetry (SD 40.7 $\mu$m, 315 $\mu$m to 579 $\mu$m).
Kaplan-Meier survival curves were generated using six different methods to define progression (Figure 5.3). The drop in survival at one year likely reflects the predefined follow-up interval of one year. For the dual parameter adaptive method, there were 1451 eyes at risk at two years after the base-line visit, 983 eyes at risk at three years, 674 eyes at risk at four years, 307 eyes at risk at five years and 84 eyes at risk at six years. The proportion of eyes that had progressed after the reference visit was 8% at two years, 12% by the end of three years, 15% by the end of four years, and 20% by the end of five years. When the fixed definition was used of a $\geq 1\text{D}$ increase in $K_{\text{max}}$, the proportion of eyes that had progressed at five years increased to 29%. In contrast, with a fixed definition of progression of a $\geq 1.5\text{D}$ increase in $K_{2}$, or the dual parameter variable definition of progression, the proportion that had progressed was lower at 8% and 6%, respectively, at five years after the baseline visit. Figure 5.4 depicts the confidence intervals for each of the methods. Depending on the definition adopted, keratoconus had progressed in 6% to 29% of eyes in the five years following the baseline visit. However, there was a marked discrepancy between keratometric progression and repeat CXL as, assuming cases did not have treatment elsewhere, only 8 eyes (0.48%) from the initial cohort of 1677 eyes had repeat cross-linking at five years after the baseline visit.

McNemar’s test was performed to compare the three adaptive progression definitions using eye status at five years. As seen in figure 5.3, the two closest lines together are Adaptive $K_{2}$ and Adaptive $K_{\text{max}}$, and under McNemar’s test to compare these two definitions, the associated p-value was very large ($P=0.947$), thus it could not be concluded they were significantly different. However, for the adaptive multi-parameter vs adaptive $K_{2}$, which are further apart in figure 5.3, the associated p-value was very small ($P<0.001$), thus they were significantly different.
Figure 5.4: The survival plots with their confidence intervals of the two methods used to calculate progression. There is an overlap of the confidence intervals for the three adaptive thresholds (left), but less so for the static thresholds, and not for the survival for the fixed threshold for Kmax (right). This suggests that there is not a statistically significant difference between the survival rates for the three adaptive threshold methods, that the survival rate calculated using the fixed threshold for Front K2 and the dual variable threshold are similar, but that both are significantly different from the fixed Kmax threshold.

5.5 Discussion

The accuracy of keratometry in keratoconus reduces as the cornea becomes more ectatic, such that a single value to define progression is not appropriate for the full spectrum of disease severity [6, 148, 5, 7]. Some studies have accounted for this variability by adopting a second, higher threshold to define progression in more advanced disease [4, 5]. To improve on this step-wise change in the definition of progression, this chapter used a 95% LoA to derive a description of progression that was applied continuously for all values of corneal curvature. This adaptive threshold defines the test-retest repeatability for each individual. Adaptive thresholds were then used to determine progression in a population with a range of keratometry values. Finally, these results were compared with the progression rate identified by standard definitions [136, 135, 134, 146, 149]. The results show that keratoconus
can continue to progress for up to six years following CXL, but with a wide range (6% to 29%) in the estimates depending on the definition that is adopted. It has also been shown that, in this population, progression does not equate to re-treatment, with only a small minority (0.46%) undergoing repeat CXL.

The effect of using adaptive thresholds to determine progression has clear implications for clinical decision making. The adaptive threshold (95% LoA, red line in figure 5.2) crosses the upper static threshold of Kmax at 53.7D and of front K2 at 64.3D. Thus below this value, the static thresholds (blue line in figure 5.2) would underestimate the number of eyes that had progressed, and above this they would overestimate progression. Therefore, the adaptive thresholds should identify the progressors in early disease, who may not have progressed according to established thresholds (e.g. 1D increase in Kmax), but have progressed beyond 95% LoA for the specific keratometry value.

When the adaptive thresholds and standard thresholds were applied to a dataset of eyes for five years following a reference examination 9-15 months following CXL, the differences in outcome becomes evident (figure 5.3). However, it is relevant that the results for the three adaptive threshold methods are similar, with overlapping confidence intervals (figure 5.4), with a five year rate of progression of approximately 20%. Unfortunately, it is difficult to compare these progression rates with other published reports due to the varied thresholds that have been used, the lack of a defined reference time point and small sample sizes although values range from 2% to 20% [150, 151, 152, 153]. This estimate of progression at two years following CXL based on the adaptive threshold multi parameter method (4%) is higher than was previously reported using the dual threshold multi parameter method (1.9 - 2.4%). The adoption of a reference interval of 9 to 15 months for the baseline visit for determining progression may also have contributed to this difference as in the earlier report which compared keratometry to pre-treatment values [135]. This delayed and longer reference interval was chosen to avoid the immediate changes in keratometry that can follow CXL, whilst also recognising that this study is conducted with real world data and that patients were unlikely to have been followed
up precisely at a defined time point. There were also marked differences between the rate of progression when defined with the single threshold Kmax compared to single threshold K2, although the magnitude of the thresholds differed (1.0D vs 1.5D). If the single threshold Kmax definition is modified to use a 1.5D as the cutoff, the 5-year progression drops from 29% to 18%, although this is still greater than the progression of 8% with a similar 1.5D threshold for K2.

In this chapter, the CXL re-treatment rate (0.46%) was markedly lower than the estimate of keratometric progression using the adaptive threshold multi parameter method at five years (14%), and lower than the more conservative dual threshold multi parameter method (3%). There are only a limited number of reports on the rate of retreatment, but there are reports of higher retreatment rates of up to 45 in 230 eyes (20%) in patients with a minimum of 36 months follow-up [149]. Of note, patients who had retreatment in this study had an average Kmax of 67.9D [149]. Using the adaptive Kmax threshold, the 95% LoA exceeds 1D when Kmax ≥ 53D, and the 95% LoA when Kmax is 67.9D corresponds to [-2.01D,1.99D]. Therefore, with the adaptive thresholds, patients with a Kmax of 67.9D would need to demonstrate more than 2 dioptres of Kmax increase for progression to be confirmed [149]. The reasons for the relatively low rate of retreatment in this cohort is uncertain, but may include a proportion of patients in whom the cornea was thinned to a value below the minimum stromal thickness considered to be safe for CXL, or a high proportion of patients who declined what was a painful experience.

The strengths of this study include the large sample size, with data recorded at each visit on a standardised spreadsheet for an extended follow-up. A threshold was incorporated for the onset of follow-up to reduce the risk of artefact from acute changes in shape unrelated to progression that follow CXL. Data was used containing repeatability limits across the spectrum of keratoconus disease severity. For the first time, the effect of using the various parameters to define progression was quantified. The study’s limitations are incomplete data and follow-up inherent in real-world databases. Older children and adults were also included, although there may be differences in their risk of progression. The survival curves were based on available
data, with a significant proportion of patients lost to follow-up; it is not known whether the characteristics of the population censored from the survival curves are the same as the included data. Finally, this is not a controlled study, and an unknown proportion of the eyes would not have continued to progress even if CXL had not been performed [134]. The Kaplan-Meier curves will thus underestimate the risk of CXL treatment failure.

In conclusion, it has been confirmed that the parameter used and the chosen threshold to define progression will affect estimates of the efficacy of CXL. Adopting personalised progression thresholds should facilitate more accurate identification of actual keratoconus progression, particularly in early disease when there are smaller changes in keratometry. For practical implementation, a guide was created that gives the clinician an estimate of the LoA for each dioptre in the studied keratometric measurements. Adaptive thresholds should be further evaluated as a consistent and personalised method to identify keratoconus progression.
Chapter 6

Discussion

6.1 Detection

This thesis has dealt with two key topics within keratoconus research: detection and progression. In chapter 2, the recent research on machine learning related to subclinical keratoconus detection was examined. This comprehensive review has details of only the most relevant algorithms focusing on subclinical keratoconus. Furthermore, the case for examining raw images as opposed to derived images was presented. In chapter 3, this idea was taken forward by training a CNN to classify early versus late stage keratoconus. Pre-segmentation of the cornea was shown to enhance the classification accuracy from 93.5% to 96.5%. Although early versus late stage classification can already be achieved without the need for deep learning, it remains an important finding because it opens the door for future research focused on classifying subclinical disease.

Subclinical keratoconus detection is the most clinically useful problem to solve. Unfortunately, finding such a dataset to train a classifier on is challenging because the labelling of such a dataset is extremely time-consuming. AS-OCT images were chosen in this research because of their availability but there exist other devices which produce suitable tomographic imaging data. In particular, MEH has a large database of Pentacam images but, to date, there is no way to bulk extract the raw Scheimpflug images due to licensing issues.
If the work in chapter 3 was taken forward to detect subclinical keratoconus and properly validated, it could be used for screening the population at risk (e.g. ages 10 to 30 years, particularly those with an atopic history and higher risk ethnicities). Artificial intelligence assistive technologies have been deployed elsewhere in ophthalmology, such as the Food and Drug Administration (FDA) approved diabetic retinopathy screening deep-learning system [154] and the deep learning system for identification of retinopathy of prematurity that has achieved FDA breakthrough status [155]. The ability to identify individuals who are at high risk of developing keratoconus would be a significant advance in clinical care with public health implications given the existence of a cost-effective low-risk treatment (CXL). In diabetic retinopathy screening, which is well established, health economic analysis has demonstrated a semi-automated model is more cost-effective compared to human graders alone, but as yet no economic studies on a purely automated system has been published [94].

The purpose of a machine learning system to detect subclinical keratoconus is to identify disease at a stage when treatment can prevent sight loss. Minimising the rate of underdiagnosis (false negatives) is therefore paramount, but the potential for harm from overdiagnosis (false positives) should not be overlooked, and the expense from over-investigation and treatment is a widespread problem in modern healthcare [156]. It is also important to quantify the patient benefit associated with the adoption of any new investigation, which cannot be ignored. For example, twenty years after FDA approval for computer aided detection (CAD) systems for mammography screening, a follow up study concluded that the CAD system can miss cancers while insurers pay more for CAD than for the prior established methods with no benefit to patients [157]. Whilst large randomised clinical trials to address these questions are not always justified or feasible for some diseases [158], in the case of subclinical keratoconus, it is eminently possible.
6.2 Progression

In chapters 4 and 5, the problem of keratoconus progression was investigated. In chapter 4, a model was developed to predict the probability of needing CXL treatment given a set of baseline corneal parameters, using advanced techniques in survival analysis. It has been translated into a web tool capable of plotting personalised projection curves. The model is the first of its kind and has direct clinical utility; it can assist with treatment planning and informing patients of their likely progression. Furthermore, although no significant SNP covariates were found, this was the first keratoconus progression model to examine genetic data. In order to translate this work into routine clinical care, the model must be validated on external data. Further work could involve applying different longitudinal data analysis techniques or using deep learning methods such as Recurrent Neural Networks (RNNs) to model progression by incorporating imaging data.

In chapter 5, a new method was devised for defining progression post CXL treatment by using a variation of the Bland Altman method to produce adaptive thresholds. This is novel because existing definitions are based on fixed thresholds which do not account for the heteroscedastic nature of corneal measurements. Various extensions are possible in this line of research such as proportional hazards analysis to find important risk factors or using non-linear methods to find the limits of agreement. It would be particularly useful to be able to assess the accuracy of a post CXL progression definition because currently there is no gold standard.

One limitation across both chapters was the quality of the dataset. It contained missing data and inconsistent appointment times. Procuring a dataset without these problems by designing a prospective research study would improve the analysis in several ways. In chapter 4, if there was no missing data, not only could it improve the model performance by expanding the number of observations, but other well-known covariates such as ethnicity or allergies could be included in the model. In chapter 5, if appointment times were consistent, then the baseline measurement could be accurately taken as exactly one year after CXL rather than an approximate band of 9-15 months. Furthermore, many eyes were discarded because either they
did not have any appointments in this period or they only had one appointment.

### 6.3 Future Directions

Before an algorithm can be developed into software as a medical device (SaMD), it needs to be developed according to regulatory processes such as ISO 13485, which require detailed documentation, risk assessment and auditing of the software development process. Furthermore, to receive approval from regulatory medical agencies, such as FDA in the US, EMA in the EU or MHRA in the UK, safety and efficacy needs to be thoroughly assessed [159]. The first phase can be achieved in a randomised case-control trial in a controlled clinical environment to assess safety, the software does not lead to dangerous decisions for the patient, and its efficacy, clinical pathways that include the software lead to improved outcomes over current practice. Further to this, health technology assessment studies can be conducted to assess the generalisability and utility of the software in a range of clinical environments (e.g. specialist hospital or optometry practices), how it influences decision-making and enhances the clinical pathway toward monitoring of disease or treatment (e.g CXL), and whether it is economically viable. When sufficient evidence of safety and efficacy has been generated and if the software fulfils the conformity assessment, it may obtain SaMD accreditation and CE marking. In the UK, following expert consultation and review, approval by NICE can lead to widespread adoption within the national healthcare setting [158, 154]. Currently, to the best of the author’s knowledge, no detection algorithm for subclinical keratoconus has been taken beyond a proof of concept stage, demonstrating diagnostic efficacy [160, 159].

To be useful it is essential that an algorithm for detection or progression can generalise beyond the limited dataset from which the model was developed and benchmarked, which requires external validation in out-of-sample datasets [161]. The creation of a large international open-source dataset of keratoconus images would help as a reference standard [162] to develop a benchmark for external validation. When generalising to external datasets the source and quality of the data should be
considered. Data from a referral hospital may not represent the general population who might be the target for screening programs, with an under-representation of eyes with mild disease.

Beyond detection and progression, genetic aetiology was briefly investigated in chapter 4 when trying to find relevant SNPs within the CXL prediction model. The influence of genetics may be further explored by defining segmented imaging features or progression parameters as endophenotypes and performing additional genome-wide association studies in order to better understand the genetic aetiology of keratoconus.

When considering the translational likelihood of this research, the methods employed in the progression chapters to analyse longitudinal data are familiar to clinicians. Therefore, it is more likely this work will find its way into clinical practice sooner than deep learning based research due to the ability of clinicians to engage with the research process. Deep learning is a relatively new field and the models are intrinsically opaque, thus research such as that found within the classification chapter will take more time for adoption.
Bibliography


[6] Elke O Kreps, Marta Jimenez-Garcia, Ikram Issarti, Ilse Claerhout, Carina Kop-


[104] Yi Xie, Lanqin Zhao, Xiaonan Yang, Xiaohang Wu, Yahan Yang, Xiaoman Huang, Fang Liu, Jiping Xu, Limian Lin, Haiqin Lin, Qiting Feng, Haotian Lin, and Quan Liu. Screening candidates for refractive surgery with corneal Tomographic-Based deep learning. 2020.


[118] Tramx. patch2image: Training FCNNs from patches to full-sized images. a framework to train arbitrarily designed networks for medical image segmentation.


[151] Marco Lombardo, Sebastiano Serrao, Giuseppe Lombardo, and Domenico


