

## DOCTOR OF PHILOSOPHY

### Statistical Modelling in Diabetic Retinopathy Research

**An in-depth aetiological analysis and predictive modelling study to enhance diabetic retinopathy detection and reduce global health inequalities in people with type 2 diabetes**

Gurudas, Sarega

*Award date:*  
2024

*Awarding institution:*  
Coventry University

[Link to publication](#)

#### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of this thesis for personal non-commercial research or study
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission from the copyright holder(s)
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

#### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# **Statistical Modelling in Diabetic** **Retinopathy Research**

*An in-depth aetiological analysis and predictive modelling study to enhance diabetic retinopathy detection and reduce global health inequalities in people with type 2 diabetes*



By:

**Sarega Gurudas**

**PhD**

**November 2023**

# **Statistical Modelling in Diabetic** **Retinopathy Research**

*An in-depth aetiological analysis and predictive modelling study to enhance diabetic retinopathy detection and reduce global health inequalities in people with type 2 diabetes*

By:

**Sarega Gurudas**

***Critical Overview Document: a Portfolio of Published Articles submitted to the Department of Intelligent Healthcare, Coventry University, in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD).***

**November 2023**



## Abstract

Public healthcare systems need innovative, cost-effective, and accessible approaches to identify people with diabetic retinopathy (DR) because the cost and training to deliver current retinal screening services to detect DR are major barriers especially in low-and-middle-income-countries (LMICs) or resource restricted settings. A risk-based approach could be a viable alternative to universal screening, however, the lack of convincing evidence on the ability of models to guide risk stratification in sight threatening diabetic retinopathy (STDR) poses a challenge for the implementation of risk-based screening strategies in LMICs. Research projects that span across developed and developing countries provide opportunities for global translation. The ORNATE-India project was designed to develop and test diverse translatable health strategies in India and the United Kingdom to tackle the global burden of diabetes-related vision impairment (VI) and reduce health inequalities among people with diabetes (PwD).

This thesis comprises of eight publications undertaken as part of the ORNATE-India project. The objectives of my thesis were to critically evaluate the statistical methods in each publication that ranged from traditional statistics to cutting-edge machine learning, with distinct aims of addressing aetiological research as well as risk prediction for DR. All statistical models employed likelihood concepts from a frequentist inferential framework. **Chapter 1** reviewed the global prevalence of DR (2015-2019), critiquing screening methods, the lack of studies in LMICs and the global differences in DR prevalence. **Chapter 2** (with 2 publications) explored the burden of VI and blindness in PwD. Publication 2 estimated the national prevalence of VI and blindness in PwD in India using survey weighted methods, showing a higher prevalence in the lower socioeconomic strata. Publication 3 considered the age-standardised incidence of VI over 10 years in proliferative DR (PDR) patients undergoing Panretinal Photocoagulation (PRP) based on direct standardization methods, highlighting the need for prompt diagnosis and treatment of PDR. **Chapter 3** explored the UK ethnic disparities in DR and STDR incidence using Cox proportional hazards models. Higher risk of DR and STDR in Black and South Asians were observed compared to their White counterparts. Moreover, risk factors of kidney function decline in PwD were found to be similar to those for STDR, substantiating the need for holistic approaches to prevention of microvascular complications. **Chapter 4** investigated diagnostic biomarkers of DR and STDR using various statistical weighting procedures. In publication 5, an environmental wide association study (EWAS) conducted on the National Health and Nutrition Examination Survey (NHANES) datasets, highlighted hyperglycemia as a key factor in DR and STDR. While publication 6, collected primary data from UK and Indian participants to assess 13 blood biomarkers for STDR screening. Cystatin-C, not collected in publication 5, emerged as a top

biomarker in comparison to those investigated, emphasising the association between renal disease and STDR. While diagnostic tools can be used to identify existing STDR, prognostic models are crucial for prevention, improving long-term health outcomes and reducing treatment costs. **Chapter 5** presents a universally applicable STDR risk tool, with model coefficients robustly verified using cox models and under the assumption of interval-censoring, demonstrating strong performance in internal (c-statistics ranging 0.778-0.832) and external validation (c-statistics ranging 0.685-0.823). The tool overcame a major barrier for implementation as it required no blood tests or technical examinations. **Chapter 6** demonstrates the development and validation of resource-driven chronic kidney disease (CKD) risk models for PwD, using fractional polynomials to model non-linear relationships and novel decision curve analysis to assess utility. These models stratify individuals based on the availability of tests, with the least invasive model eliminating the need for blood tests or technical examinations other than kidney markers eGFR and albumin to creatinine ratio (ACR). Finally, **chapter 7** summarises the entire doctoral work, evaluating the ORNATE-India project's impact and my contributions that led to its success.

## Acknowledgements

I would like to acknowledge the following people for their support and guidance throughout my PhD:

- Prof Robyn Tapp, who graciously accepted the role of my director of studies, has been a constant source of guidance, mentorship, and moral support throughout the last 2 years. Her expertise in diabetes and DR has greatly enriched my work, and I appreciate the time and effort she dedicated to our regular meetings and collaborations since the beginning of my PhD journey.
- Prof Toby Prevost, who provided extensive support as a senior statistician throughout the Ornate India project, contributing significantly to my growth as a statistician. His valuable insights have been influential in my development.
- Prof Sobha Sivaprasad, for her instrumental role in launching my research career. Most notably, recruiting me onto the Ornate India project, helping me get to grips with the world of ophthalmology research, connecting me with Prof. Tapp, and for her encouragement in undertaking this PhD. Her encouragement has been a driving force in my academic journey.
- To the clinical fellows at Moorfields Eye Hospital (MEH) who have taught me so much about the field, and for their patience and time.
- To all my colleagues on the ORNATE-India project for their companionship and support.
- To my colleagues who collected the data; field workers, laboratory scientists and researchers, clinicians and graders who graded and processed countless retinal images.
- To the funders; the United Kingdom Research and Innovation Global Challenges Research Fund (UKRI-GCRF) and National Institute of Health and Care Research Moorfields Biomedical Research Centre (Moorfields NIHR BRC) for funding my role.
- To the patients who provided consent for their data to be used.
- Finally, to my family for their patience and support throughout my journey.

# Table of contents

<b>Abstract .....</b>	<b>3</b>
<b>Acknowledgements .....</b>	<b>5</b>
<b>List of figures .....</b>	<b>8</b>
<b>List of tables.....</b>	<b>9</b>
<b>Abbreviations.....</b>	<b>10</b>
<b>Note to the Reader .....</b>	<b>13</b>
<b>Background and Rationale.....</b>	<b>14</b>
Diabetes Mellitus and its Global Challenges .....	14
Epidemiology and Global Challenges of DR .....	14
Risk Factors and Associations between DR and DKD .....	16
<b>Portfolio and Objectives .....</b>	<b>18</b>
<b>Autobiographical context of the portfolio .....</b>	<b>21</b>
<b>Chapter 1. Global Burden of DR.....</b>	<b>23</b>
Context and Objective .....	23
Methodological commentary and Critical powers .....	23
Results .....	24
Originality and Contribution to the subject .....	24
Critical reflection.....	26
<b>Chapter 2. Prevalence and Incidence of VI in India .....</b>	<b>27</b>
Context and Objective .....	27
Methodological commentary and Critical powers .....	27
Results .....	31
Originality and Contribution to the subject .....	34
Critical reflection.....	35
<b>Chapter 3. Ethnic differences in DR .....</b>	<b>37</b>
Context and Objective .....	37
Methodological commentary and Critical powers .....	37
Results .....	39
Originality and Contribution to the subject .....	41
Critical reflection.....	41
<b>Chapter 4. Diagnostics tests for DR .....</b>	<b>43</b>
Context and Objective .....	43
Methodological commentary and Critical powers .....	44
Results .....	45

Originality and Contribution to the subject .....	50
Critical reflection.....	51
<b>Chapter 5. Prognostic Modelling in Sight Threatening Diabetic Retinopathy .....</b>	<b>53</b>
Context and Objective .....	53
Methodological commentary and Critical powers.....	53
Results .....	55
Originality and Contribution to the subject .....	60
Critical reflection.....	60
<b>Chapter 6. Prognostic modelling in CKD using routine EHR data .....</b>	<b>61</b>
Context and Objective .....	61
Methodological commentary and Critical powers.....	61
Results .....	63
Originality and Contribution to the subject .....	66
Critical reflection.....	66
<b>Chapter 7. Synthesis.....</b>	<b>68</b>
Contextualisation of research, Impact, and Study strengths .....	68
Limitations .....	71
Contribution to the field, Implementation challenges and Future work .....	72
Contributor statements .....	74
Development and growth as a researcher .....	75
Conclusion .....	79



## List of figures

Figure 1. Publications flow diagram.....	20
Figure 2. Forest plots showing the prevalence of any DR in T2DM by 7 IDF regions .....	25
Figure 3. The clinical centres and sites participating in the SMART-India study .....	28
Figure 4. Field workers in SMART-India study .....	30
Figure 5. Figure Odds Ratio plot showing risk factor burden for VI in known and undiagnosed diabetes from adjusted survey weighted logistic regression .....	32
Figure 6. Illustrating the risk of DR and by ethnicity in prevalent T2DM at baseline .....	40
Figure 7. Diagnostic performance of biomarkers for STDR selected from a forward stepwise routine with AUC-ROC curves and 95% CIs for UK and India .....	50
Figure 8. Calibration plots for model's 1, 2 and 3 showing observed vs predicted 3-year risk of STDR in validation cohorts.....	58
Figure 9. Risk-chart showing 3-year % risk of STDR using the non-invasive model (Model 3) in model development cohort.....	59
Figure 10. Decision curves comparing CKD models in the external validation cohort.....	64
Figure 11. Risk score interpretation for the prediction of 5-year risk of stage 3 + CKD (minimal resources model) .....	65
Figure 12. Contributorship matrix showing my contributions for P1-P8.....	75
Figure 13. Citations received in each year based on P1-P8, updated 16/03/2024.....	76
Figure 14. Network model for research connections formed based on P1-P8 .....	76

## List of tables

Table 1. List of publications included in this thesis – title, link to publication and references. ....	18
Table 2. Chapter 1 Publication 1 with citation and mentions, updated 16/03/2024. ....	23
Table 3. Chapter 2 Publication 2 and 3 with citations and mentions, updated 16/03/2024. ....	27
Table 4. Estimated prevalence of VI based on US and WHO severity scale .....	31
Table 5. Age stratified ten-year crude incidence of Visual Impairment and Blindness based on best corrected visual acuity .....	33
Table 6. Chapter 3 Publication 4 with citations and mentions, updated 16/03/2024. ....	37
Table 7. EHR data quality checks .....	38
Table 8. Chapter 4 publications 5 and 6 with citations and mentions, updated 16/03/2024. ....	43
Table 9. Statistically significant laboratory variables following FDR correction, associated with DR in PwD .....	47
Table 10. Chapter 5 Publication 7 citations and mentions, updated 16/03/2024. ....	53
Table 11. Hazard ratios for risk models predicting three-year risk of STDR. ....	56
Table 12. Chapter 6 Publication 8, citations and mentions, updated 16/03/2024. ....	61
Table 13. Altmetric and Plumx metrics for P1-P8, updated 16/03/2024. ....	77

## Abbreviations

ACR	Albumin to Creatinine Ratio
ASHA	Accredited Social Health Activist
AUC	Area Under the Curve
BCVA	Best Corrected Visual Acuity
CKD	Chronic Kidney Disease
COVID-19	Coronavirus Disease-19
CPRD	Clinical Practice Research Datalink
CRP	C-Reactive Protein
DESP	Diabetic Eye Screening Programme
DKD	Diabetic Kidney Disease
DM	Diabetes Mellitus
DMO	Diabetic Macular Oedema
DR	Diabetic Retinopathy
EHR	Electronic Health Record
EM	Expectation-Maximisation
ESRD	End-Stage Renal Disease
ETL	Epidemiological Transition Level
EWAS	Environment Wide Association Study
FDR	False Discovery Rate
GCRF	Global Challenges Research Fund
GP	General Practice
GWAS	Genome-Wide Association Study
HbA1c	Glycated Hemoglobin
HDL	High Density Lipoprotein
HIC	High Income Country
IDF	International Diabetes Federation
KDIGO	The Kidney Disease: Improving Global Outcomes group
LMIC	Low-and-Middle Income Country
logMAR	Logarithm of the Minimal Angle of Resolution
MAR	Missing At Random
MDRF	Madras Diabetes Research Foundation

MEH	Moorfields Eye Hospital
META-EYE	Meta-analysis for Eye Disease
NHANES	National Health and Nutrition Examination Survey
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NPDR	Non-Proliferative Diabetic Retinopathy
NPMLE	Nonparametric Maximum Likelihood Estimation
OR	Odds Ratio
PCA	Principle Component Analysis
PDR	Proliferative Diabetic Retinopathy
POC	Point-Of-Care
PROGRESS	The PROGnosis RESearch Strategy
PRP	Pan-Retinal Photocoagulation
PwD	People with Diabetes
RAAB	Rapid Assessment of Avoidable Blindness
RF	Random Forest
ROC	Receiver Operating Characteristic Curves
SAIL	Secure Anonymised Information Linkage
SMART-India	Translating research into clinical and community practice: a multicenter Statistical and economical Modelling of risk-based strAtified and peRsonalised screening for complications of diabetes in India
SN-DREAMS	Sankara Nethralaya-Diabetic Retinopathy Epidemiology And Molecular genetic Study
STDR	Sight Threatening Diabetic Retinopathy
STROBE	Strengthening The Reporting of Observational studies in Epidemiology
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
TRIPOD	Transparent Reporting of multivariable prediction model for Individual Prognosis Or Diagnosis
UCL	University College London
UKPDS	UK Prospective Diabetes Study
UKRI	UK Research and Innovation
VA	Visual Acuity
VI	Vision Impairment
WHO	World Health Organisation

WP	Work Package
----	--------------

## Note to the Reader

This critical overview document serves as a synthesis of eight publications, with all publications being open access except for P1. Relevant references and links to the publications can be found in the “Portfolio and Objectives” section and it is recommended that they are reviewed before or in conjunction with their respective chapters. P1 has been appended to this thesis for reference.

While a PhD thesis typically employs a third person perspective, the requirements for a PhD by publication necessitate the inclusion of critical reflection sections throughout where I discuss my contributions and progression as a researcher, which will be authored in the first person.

# Background and Rationale

## Diabetes Mellitus and its Global Challenges

Currently Diabetes Mellitus (DM) impacts 537 million adults (20-79 years) globally, with a projected increase to 783 million adults by 2045 [1]. Type 1 diabetes mellitus (T1DM) results from insulin deficiency while in Type 2 diabetes mellitus (T2DM), accounting for over 90% of DM cases, the body produces insulin but is not used effectively [2, 3]. Sub-optimally controlled diabetes may lead to microvascular (small blood vessel) complications, impacting the retina (DR), kidneys (Diabetic kidney disease (DKD)) and the nervous system (including Diabetic peripheral neuropathy), and/or macrovascular (large blood vessel) complications, causing stroke and heart disease.

LMICs are most affected by the diabetes epidemic, with India boasting one of the largest populations of adults with diabetes. An estimated 101 million people in India had diabetes in 2021 [4] and this is expected to rise to 134.2 million by 2045 [5]. It is a global health priority aligned with the United Nations Sustainable Development Goal 3 to reduce diabetes and its complications [6]. The UK's National Health Service (NHS) has a well-established diabetes care pathway, but considering limited resources, reducing costs for existing programs remains a top priority [7]. Approximately, £10 billion is allocated annually for diabetes care, accounting for 10% of the NHS budget [8, 9].

## Epidemiology and Global Challenges of DR

DR, one of the most prevalent microvascular complications of diabetes and leading cause of avoidable blindness in the working-age group [10], can be categorised into three stages; i) background retinopathy (mild / moderate non-proliferative DR (NPDR)), ii) pre-proliferative (severe NPDR) and iii) PDR [11]. Vision becomes affected due to complications of PDR or diabetic macular oedema (DMO). Two of the most common treatments for DR include intravitreal injections or PRP in STDR. The Diabetic Retinopathy Study Research Group (1976) reported that PRP led to a 57% reduction in the risk of vision loss in those with proliferative DR [12], and more recent studies such as protocol S showed PRP led to stable vision in 60% of active PDR within 5-years [13]. However, side effects include; reduced visual field and central vision, night-blindness, vitreous hemorrhage and macular oedema [14-16].

To date, studies have reported the diagnosis of DR largely using ophthalmoscopy and retinal photography [17]. However, the former fails to meet the British Diabetic Association's (Diabetes UK) and UK National Institute for Clinical Excellence (NICE) 80% sensitivity and 95% specificity targets [18, 19]. Therefore, there is a need to update the global prevalence based on standard retinal photographs to avoid ascertainment bias.

A 2021 meta-analysis reported the global prevalence of DR and STDR in PwD to be 22.27% and 6.17% respectively [20]. In India, estimates from the SMART-India study reported prevalence as 12.5% for DR, and 4.0% for STDR [21]. In the UK, it is 32.10% for DR and 10.99% for STDR [22]. Although the UK has a higher proportion of DR and a rising diabetes prevalence, systematic DR screening has prevented it from being the leading cause of blindness in working-age adults [23]. International guidelines recommend that PwD should undergo annual digital retinal photography as evidence has shown that early detection and timely treatment can reduce vision loss by about 95% [11, 24, 25]. However retinal screening is costly and not universally available [26]. Current DR screening in the DESP relies on costly two-field 45° color fundus photography [27]. Recent studies suggest that individualised risk-based screening intervals could alleviate strain on resources due to annual retinal photography, given the rising burden of diabetes and limited budgets. In fact, in 2016, the UK National Screening Committee (NSC) recommended extending the screening interval for individuals at low risk of DR from annual to biennial screenings. The shift was based on a large study that demonstrated a low incidence and progression of DR over a two-year period in low-risk eyes [28]. However, low-risk eyes were determined based on stage of retinopathy without considering other systemic risk factors. For strategies to be feasible in both developed and developing countries, further research is needed to develop alternative screening strategies that don't rely on retinal imaging.

In LMICs like India, retinal screening is still in its infancy, resulting in a lack of data on the prevalence of VI and blindness in high-risk groups like PDR, and studies with national coverage. Preventing or delaying the onset of DR through optimal risk factor control is key in managing the burden of eventual blindness due to STDR. These risk factors include suboptimal glycaemic control, hypercholesterolemia, presence of hypertension, ethnicity and longer duration of diabetes [9, 29]. Data on the incidence of DR and STDR among ethnic minority populations in the UK is limited and much needed to plan prevention programs. Several other blood parameters are associated with diabetes but the significance of these parameters in DR remains to be explored [9, 30].



## Risk Factors and Associations between DR and DKD

As with other diabetes complications, prevalence of DKD varies by country, with LMICs seeing a prevalence of around 15% in people with newly diagnosed T2DM [31] and in other studies the prevalence in T2DM has varied from 27%-87% [32-37]. DKD is the most frequent cause of end stage renal disease (ESRD) [38, 39] and ranks 12<sup>th</sup> in global causes of death [40]. It's classified by combining GFR stages (G1-G5) and albuminuria categories (A1-A3) [41], with early stages often asymptomatic necessitating regular monitoring for timely intervention. For those who have progressed to kidney failure, treatments like dialysis and kidney transplantation are involved, with dialysis impacting quality of life [42]. The kidney disease: improving global outcomes (KDIGO) group, established in 2003, provides clinical guidelines to enhance DKD detection and management in PwD [43].

DR and DKD share many common risk factors that are modifiable (i.e., poor blood glucose control, lipids, hypertension, smoking) and non-modifiable (longer duration of diabetes, age, gender and ethnicity) [44, 45]. This indicates the importance of comprehensive screening when one complication is present [46]. Moreover, studies have found retinal vascular signs to be associated with presence of DKD [47-50]. Therefore, prevention efforts can be optimised to allow for a more holistic approach to diabetes care. Previous prevention efforts have been largely focused on glycemic control [51]. HbA1c has long served as an established marker for diabetes complications and Cystatin-C has also emerged as a robust indicator for DR reflecting the DKD-DR relationship. With clinical biomarker testing requiring samples to be analysed by trained individuals, high operating cost, long waiting lists and lack of service access for those residing in remote areas of India intensifies the need for point-of-care (POC) devices. While many circulating markers have also been shown to be distinguished in DR, there is a need to evaluate these markers in larger cohorts [30, 52, 53].

Multiple-marker models may be more specific in detecting DR, as several studies have shown [52, 53], however optimisation of multiplex detection may be required before integration into a POC biosensor. Research evaluating diverse biomarkers for DR employing data-driven statistical modeling techniques like machine learning for high-dimensional datasets and theory-driven methods involving comprehensive literature reviews are essential.

Prognostic models can serve as a tool to assist PwD in managing and preventing complications like STDR and DKD [44]. Existing studies have limitations, including high-cost, biased statistical analysis, lack of external validation, or inadequate discriminatory ability, emphasizing the need for well-designed studies, producing resource-driven models to reduce the costs associated with

population level screening [7, 44]. A recent model by RetinaRisk achieved satisfactory discriminatory ability but required laboratory and retinal imaging data [54], therefore further work is needed to improve the transportability of these models to LMICs.

The ORNATE-India project, a UK Research and Innovation (UKRI) and Global Challenges Research Fund (GCRF) funded UK-India multidisciplinary research collaboration was set up in 2019 to tackle the burden of VI due to diabetes, and address research gaps on identifying and developing resource-driven screening strategies for DR which overcome implementation challenges in LMICs [55]. My PhD research was undertaken as a part of this project.

## Portfolio and Objectives

This portfolio contains a selection of my publications focused on the development of DR detection tools, spanning the period between 2019 and 2024. This synthesis consisting of 8 publications across 6 chapters, constitutes a single and coherent narrative tracing the development of my research within the ORNATE-India project (**Table 1**). All publications utilise quantitative research methods employing diverse statistical models tailored to the available data and research question.

**Table 1. List of publications included in this thesis – title, link to publication and references.**

<u>Chapter</u>	<u>#</u>	<u>Title and link to publication</u>	<u>Ethics approval</u>	<u>Ref</u>
1	P1	<u>IDF Diabetes Atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018</u>	Not required <sup>a</sup>	[17]
2	P2	<u>National prevalence of vision impairment and blindness and associated risk factors in adults aged 40 years or older with known or undiagnosed diabetes: results from the SMART-India cross-sectional study</u>	HMSC/2018-0494 [56] <sup>b</sup>	[57]
	P3	<u>Prevalence and incidence of visual impairment in patients with proliferative diabetic retinopathy in India</u>	Not required <sup>c</sup>	[58]
3	P4	<u>Ethnic Disparities in the Development of Sight-Threatening Diabetic Retinopathy in a UK Multi-Ethnic Population with Diabetes: An Observational Cohort Study</u>	Not required <sup>c</sup>	[59]
4	P5	<u>Diabetic Retinopathy Environment-Wide Association Study (EWAS) in NHANES 2005-2008</u>	Not required <sup>d</sup>	[60]
	P6	<u>Multicenter Evaluation of Diagnostic Circulating Biomarkers to Detect Sight-Threatening Diabetic Retinopathy</u>	REC: 18/SC/0477 <sup>e</sup>	[61]
5	P7	<u>Development and validation of predictive risk models for sight threatening diabetic retinopathy in patients with type 2 diabetes to be applied as triage tools in resource limited settings</u>	Not required <sup>c</sup>	[62]
6	P8	<u>Development and validation of resource-driven risk prediction models for incident chronic kidney disease in type 2 diabetes</u>	Not required <sup>c</sup>	[63]

Abbreviations; IDF- International Diabetes Federation, NHANES- National Health And Nutrition Examination Survey, HMSC- Health Ministry Screening Committee, REC- Research Ethics Committee.<sup>a</sup> Review of the literature. <sup>b</sup> Approvals detailed in the SMART-India study protocol. <sup>c</sup>

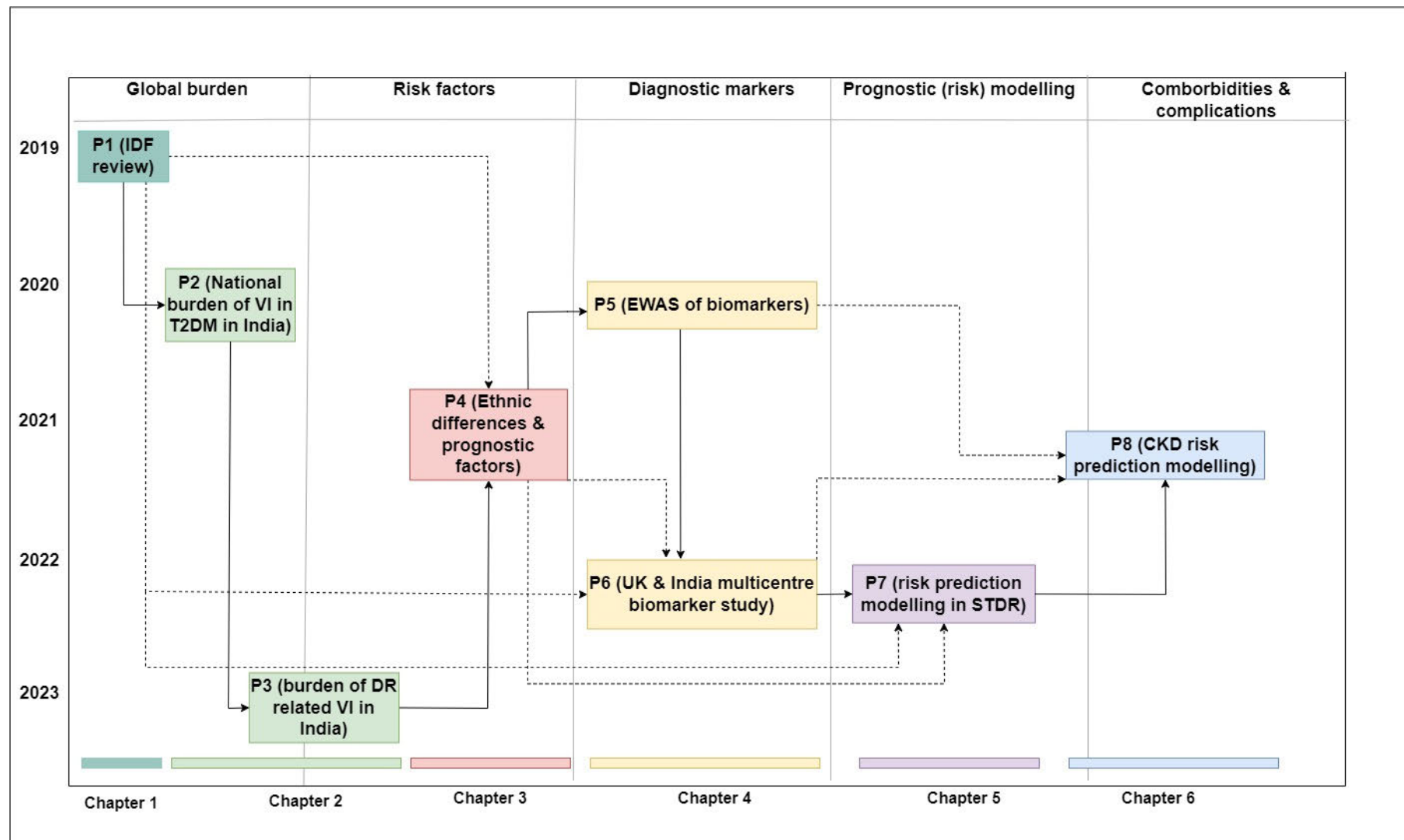
*Fully anonymized retrospective routine data.* <sup>d</sup> *Dataset available in the public domain.* <sup>e</sup> *Health Research Authority and Health and Care Research Wales (HCRW) approval (REC reference 18/SC/0477).*

This synthesis aims to present the pipeline of my research, from the global burden of DR, the links with DKD, its wider impact on health and vision, as well as the development and validation of risk models in datasets spanning from the developed and developing worlds. Objectives include:

1. To estimate the current global prevalence of DR, investigate its epidemiology and outline research gaps in the literature (**Chapter 1**)
2. To estimate the prevalence of VI in PwD in India with national coverage (**Chapter 2**)
3. To estimate the prevalence and incidence of PDR-related VI in India (**Chapter 2**)
4. To examine the ethnic differences and risk factors in DR (**Chapter 3**)
5. To explore resource-driven strategies that aid diagnosis of STDR (**Chapter 4**)
6. To develop resource-driven prognostic statistical models using routine variables to predict those at risk of progression to STDR (**Chapter 5**)
7. To examine the utility of routine variables for predicting risk of progression to DKD (**Chapter 6**)

**Figure 1** provides a visual guide of the link between the chapters and publications and the phase of research to which they align; from the global burden of DR (chapter 1 and 2), key risk factors (chapter 2 and 3), diagnostic markers (chapter 4), prognostic modelling (chapter 5 and 6) and comorbidities and complications (chapter 6). **Appendix 1** provides further details on the links between publications.

Figure 1. Publications flow diagram



## Autobiographical context of the portfolio

Prior to joining University College London (UCL) Institute of Ophthalmology (IoO) and Moorfields Eye Hospital (MEH), I pursued a master's degree (in 2017) in medical statistics at UCL's Statistical Science Department, where my journey into research began. My masters project involved Bayesian modelling to predict the gentamicin levels in newborns receiving treatment for infections. This experience ignited my passion for prediction modelling research, leading me to join the IoO research team in early 2018 as a research assistant. At IoO, I developed a pragmatic research approach, as most of our study outcomes were measured against its viability in LMICs. I worked with experts from multiple disciplines, honing my skills in big data analysis as well as in analysis of smaller datasets. I also organised and delivered statistics workshops for MSc Ophthalmology students, promoting collaborative research between ophthalmologists and statisticians.

The global DR literature review was my first publication in my research career (**chapter 1**). This experience provided me with an early insight into the project's main disease area, and the extent of the DR problem. It also helped me become acquainted with the field's terminology, the typical challenges encountered in DR research and the complexities of working with ophthalmic data from around the world. To gain a deeper understanding of its impact, I became engaged in two additional projects. First, I contributed to the VI in PDR study where I developed the revised statistical analysis plan (**chapter 2; P3**) and conducted the analysis. Subsequently, I began working on a study aimed at developing risk models for CKD in PwD (**chapter 6**). The main work packages (WP's) within ORNATE-India included WP5, focusing on risk prediction modelling. Consequently, a significant portion of my time was dedicated to planning and preparing for the analysis of primary care ("big") data. For the study presented in **chapter 5 (P7)** and **chapter 6 (P8)**, my colleagues and I drafted protocols and statistical analysis plans, which were shared with data owners of the East London General Practice (GP) dataset, Secure Anonymised Information Linkage (SAIL) databank, and Madras Diabetes Research Foundation (MDRF). I learnt how to correctly develop and validate a risk model using the latest guidance. It was also the first time I used interval-censored cox models in a dataset of that size, where I faced computational challenges. During this time, the biomarker study was underway, which had its own share of challenges, including a freezer malfunction that resulted in loss of UK blood samples for the study (**chapter 4; P6**). This setback prompted a restart to data collection, causing a delay in the analysis. However, this allowed me to focus on other research activities. Recognising the need for data-driven investigations involving multiple markers, an investigation was initiated into diagnostic markers using an existing publicly available dataset. This study served as a preamble

to the biomarker investigation in **P6**, which was already underway. In 2020, the EWAS study was published confirming several mechanisms of action in DR and validating HbA1c's utility over 400+ laboratory variables collected from NHANES (**chapter 4; P5**). This was also my first-time applying survey weights as I recognised the need to account for the survey design of NHANES. A key limitation was the inability to conclude causal relationships, highlighting the need for both short-term and long-term data. As I embarked on the risk modeling studies (**chapter 5 and 6**), I recognised the need to study risk factors in detail (**P4**). Similar risk factors were included in CKD and DR risk modelling studies, where the close relationship between the diseases became more evident in my work. Finally, in response to the publication of the SMART-India study in 2022, which emphasised the need for robust data of vision outcomes in PwD, the vision data was subsequently published in 2024 (**chapter 2; P2**).

As my time at UCL concluded, I transitioned to my current role as a medical statistician at MEH where I continued to collaborate with the same supervisory team, supporting observational studies related to medical retina. Most recently I assumed the role as a lead author on a working paper on health-related quality of life in PwD, that has been accepted for presentation at the International Diabetes Federation (IDF) world congress 2023. Over time, I have assumed increased responsibilities in my role, and in the future aim to broaden my statistical toolkit by gaining experience in clinical trials, securing independent funding and advancing to a leadership position.

# Chapter 1. Global Burden of DR

## Context and Objective

Understanding the global prevalence and epidemiology of DR was a crucial first step to uncovering global variations in DR rates and to facilitate effective healthcare planning. Two previous reviews that provided estimates on the global prevalence of DR had major limitations. A 2012 meta-analysis (META-EYE study) [64] spanning a large time frame (1980-2008) and observing a wide range of prevalence of DR on studies that assessed fundal photographs, was likely to have been biased by time period effects, meaning that heterogeneity (diversity) between studies made comparisons difficult. The review by Lee et al in 2015 comprised a more recent period (2003-2015) and adopted a much broader search strategy. Updates to the prevalence of DR covering shorter time frames are needed to reduce the impact of heterogeneity .

To account for these limitations, publication 1 (**P1**) (**Table 2**) estimated the global prevalence of DR covering studies published 2015-2019 using digital retinal photography, by 7 IDF regions (Africa, Europe, Middle East and North Africa, North America and Caribbean, South and Central America, South-East Asia and the Western Pacific), updating the 2015 review [24]. The study also investigated the global epidemiology of DR and identified research gaps in a lack of studies from the developing world. The manuscript featured in the 9<sup>th</sup> edition of the IDF Atlas [5].

**Table 2. Chapter 1 Publication 1 with citation and mentions, updated 16/03/2024.**

P1: DR literature review

IDF Diabetes Atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018

Thomas, R.L., Halim, S., **Gurudas, S.**, Sivaprasad, S., & Owens, D  
2019, DRCP, Impact factor: 8.18



## Methodological commentary and Critical powers

To minimise risk of information bias and selection bias, and to improve reviewing accuracy, 4 databases were searched (PubMed, Embase, Web of Science and Medline) with MESH terms diabetic retinopathy and prevalence, with articles hand sifted and duplicates removed manually. Secondly, there was a lack of uniformity across studies in relation to the grading criteria and retinal capturing methods, also many studies did not report whether mydriasis was used or not, therefore this information was not used to synthesise our findings. Thirdly, the review was based on studies



that underwent digital retinal photography, for uniformity and accuracy, and excluded studies which may have used direct or indirect ophthalmoscopy or any other forms of screening. Fourthly, the international clinical DR and DMO disease severity scales were used [65] to standardize the recording of presence and severity of DR and DMO, therefore results are presented as prevalence of any DR, NPDR (mild DR, moderate and severe pre-proliferative DR), PDR and DMO (DMO and/or clinically significant macular oedema).

## Results

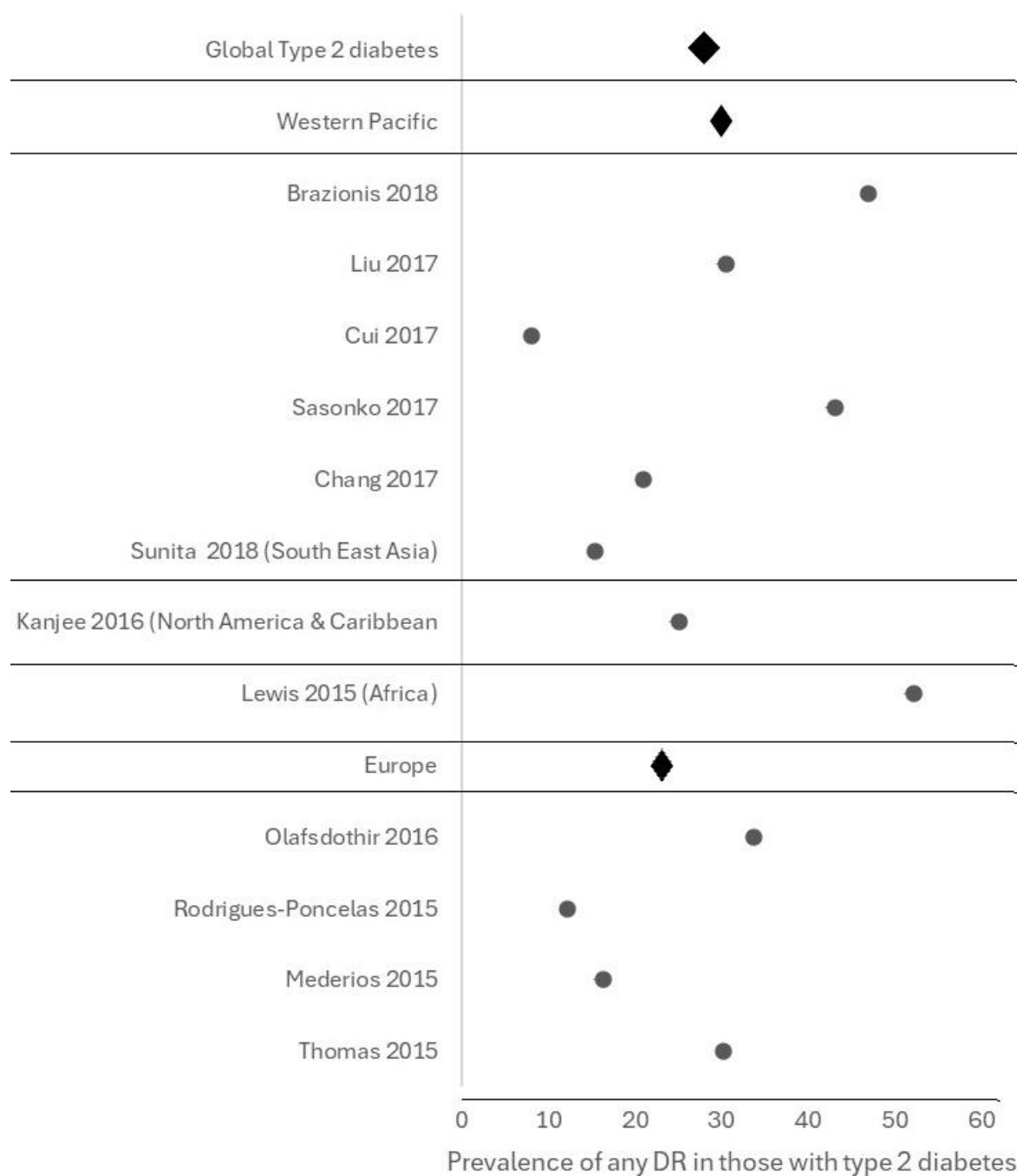
In total 32 articles were included in the review. The global prevalence of DR combining type 1 and T2DM was estimated as 27.0% for any DR (25.2%, NPDR, 1.4% PDR and 4.6% DME). The lowest prevalence was in South-East Asia at 12.5% and Europe, 20.6% and highest seen in the Western Pacific region (36.2%), Africa (33.8%) and Middle East and North Africa (33.8%). **Figure 2** shows the prevalence in those with T2DM only. The included studies lacked uniformity in the study population definitions, there were methodological variations, and differences in the DR grading criteria.

## Originality and Contribution to the subject

At the time of publication, the most recent estimates of global DR prevalence were outdated, relied on inconsistent screening methods, or covered a broad search strategy, thereby making comparisons between studies difficult. Providing the latest estimates for the prevalence of DR was also needed to plan required provision of health services to manage this global burden and reduce the rate of VI in PwD.

This paper is a valuable source of evidence towards improving current knowledge on global DR burden. Consequently, it is the most cited and widely read publication within the portfolio. The review provides a detailed breakdown on the regional, national, and global prevalence of DR in varying severity on studies that fulfilled our strict inclusion/exclusion criteria to inform local governments, policy makers and health professionals. Finally, it highlights the lack of global consensus on the screening (process of retinal image capture) and classification guidelines (grading) in DR and the need for an internationally agreed screening and diagnostic criteria to deriving the true global prevalence of DR.

**Figure 2. Forest plots showing the prevalence of any DR in T2DM by 7 IDF regions**



*Key: Summary population by region – Diamond. Abbreviations: DR- Diabetic Retinopathy. Extracted from P1 [17].*

## Critical reflection

This was my first published article and marked the beginning of my academic career. Through this study, I deepened my knowledge of current global DR literature, including the terminologies, classification of DR, and its global epidemiology. I played a key role in refining data collection processes, shaping the final publication tables. The review helped me identify several research gaps and future considerations which helped frame the objectives in subsequent publications, such as ethnic, regional disparities in DR, DR burden in blindness and the challenge of diabetes in LMICs.

I believe the study was well-conducted and sufficiently rigorous with a robust search criterion however lacked population-based studies from developing countries, however similar conclusions were drawn in the Meta-analysis for Eye Disease (META-EYE) study 2012 [64] which noted a higher prevalence of DR in T1DM than T2DM. Still, there were several limitations which may have impacted on the quality of our reported findings. Heterogeneity due to the impact of differing time periods is a somewhat lesser concern, yet they cannot be completely ignored. This study also acknowledged that the included studies did not provide good coverage within each IDF region. Consequently, the distribution of studies across some of the IDF regions were sparse, with as few as 1 study contributing to the South America and Caribbean region and 1 from Middle East and North Africa. Admittedly, due to time constraints, opportunities to undertake a systematic review with a higher level of evidence were not pursued. Due to heterogeneity between studies (differing methods of retinal capture such as the extent and location of retinal fields, whether mydriasis was used, grading criteria, lack of uniformity in study population definitions including self-reported data on type of diabetes), it was concluded that a meta-analysis by taking the weighted average of the estimates, may not yield meaningful analyses. Moreover, summary estimates should be interpreted with caution.

In summary **P1** established the global prevalence of DR, its disparities and highlighted the need for studies to be conducted in LMICs. The study the most cited publication in this portfolio (**Table 2**) with a total of 197 citations and highlighted several disparities in DR classification and method of retinal capture, where consensus is required to help improve global clinical practice and research in DR.

## Chapter 2. Prevalence and Incidence of VI in India

### Context and Objective

To gain a comprehensive understanding of the impact of DR, it was crucial to update prevalence of VI with robust, nationally representative data on PwD in India. The latest report in India used ophthalmoscopy for DR diagnosis and did not study key risk factors of VI in PwD [66]. Furthermore, a major consequence of end-stage retinopathy (PDR) is vision loss, associated with reduced quality of life [67]. Therefore, studies that focused on the prevalence and incidence of VI in DR were needed to provide a comprehensive assessment of the burden of DR.

Considering this, **P2 (Table 3)** was conceived to provide national estimates of the prevalence of VI in India for the 2018-2020 period. While **P3** focuses on vision outcomes in PDR patients undergoing PRP to better understand where resources should be targeted to maximise care and health outcomes in high-risk groups.

**Table 3. Chapter 2 Publication 2 and 3 with citations and mentions, updated 16/03/2024.**

P2: Prevalence of VI and blindness study

National prevalence of vision impairment and blindness and associated risk factors in adults aged 40 years or older with known or undiagnosed diabetes: results from the SMART-India cross-sectional study

**Gurudas S**, Joana C Vasconcelos, A Toby Prevost, Rajiv Raman, Ramachandran Rajalakshmi, Kim Ramasamy, Viswanathan Mohan, Padmaja K Rani, Taraprasad Das, Dolores Conroy, Robyn Tapp, Sobha Sivaprasad on behalf of SMART-India Study Collaborators  
2024, Lancet Global Health, Impact factor: 34.3



P3: Prevalence and incidence in PDR study

Prevalence and incidence of visual impairment in patients with proliferative diabetic retinopathy in India

Khan R, Chandra S, Rajalakshmi R, Rani PK, Anantharaman G, Sen A, Desai A, Roy R, Natarajan S, Chen L, Chawla G, Behera UC, Gopal L, **Gurudas S**, Sivaprasad S, Raman R  
2020, Scientific reports, Impact factor: 4.011



### Methodological commentary and Critical powers

**P2** was the largest primary dataset collected in the project, with approvals detailed in the study protocol [56], aiming to provide prevalence estimates with national coverage in a cross-section of

the population. While **P3** provided longitudinal data considering the trajectory to VI and blindness in PDR patients. The methodological decisions are detailed below.

## P2

Firstly, due to the complex cluster sampling design, involving stratification and clustering, study weights were derived to ensure our estimates were nationally representative. These weights were published in the original DR prevalence study [21], however the current study's analysis sample included individuals with ungradable images, necessitating the derivation of new study weights. At least one state per region was sampled, with participating sites shown in **Figure 3**. Sample weights were calculated by comparing the number of participants in each stratum to the region-specific national diabetes population in India. This population within each of the six regions was estimated by multiplying 2011 Census of India data with state-wise diabetes rates from the Global Burden of Disease study [68] and rural-urban populations based on the Indian Diabetes (INDIAB) study, which reported rates of 8.9% and 16.4% in rural and urban areas respectively [4].

### **Figure 3. The clinical centres and sites participating in the SMART-India study**

**This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.**

*Extracted from Ornate India protocol[69]*

Survey-weighted logistic regression was performed, weighting and clustering by enumeration district using the R Survey package [70]. This method involves survey-weighted maximum

likelihood estimation where the log-likelihood of the logistic regression model is adjusted by the survey weights (**Equation 1**) [71, 72]. The modified log-likelihood is known as pseudolikelihood due to the probabilistic interpretations no longer being applicable. The standard formulations' stochastic assumptions may not reflect the complex population structure inherent in the sampling, where observations for different individuals are correlated within clusters.

**Equation 1. Modified likelihood function for logistic regression with probability weights**

$$L(\beta) = \sum_{k=1}^K \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} w_{hji} [y_{hji} \log(p_{hji}) + (1 - y_{hji}) \log(1 - p_{hji})]$$

$w_{hji}$  is the survey weight for the  $hji$ -th observation

$y_{hji}$  is the binary dependent variable for the  $hji$ -th observation (0 or 1)

$p_{hji}$  is the predicted probability that  $y_{hji} = 1$  based on the logistic regression model

$K$  is the number of strata

$m_k$  is the number of primary sampling units for the  $k$ -th strata

$n_{kj}$  is the number of elements in the  $kj$ -th sampling unit

$L(\beta)$  is the log-likelihood function

Secondly, known confounders of VI and blindness including age, gender, diabetes duration, education, Epidemiological transition level (ETL) and rurality were considered. Thirdly, VI and blindness was defined based on the US criterion, as the Peek Vision application used to record visual acuity (VA), truncated values at 1.3 LogMAR and above. However, estimates for mild and moderate VI were reported based on the World Health Organisation (WHO) criterion, to aid comparability with the literature, e.g. with the Global Burden of Disease [73] or Rapid Assessment of Avoidable Blindness (RAAB) studies [66, 74]. Finally, DR was graded based on the international severity scales [65] and retinal images were captured using hand-held retinal cameras using non-mydratic screening (**Figure 4**).

**Figure 4. Field workers in SMART-India study**



**This item has been removed due to 3rd Party Copyright. The unabridged version of the thesis can be found in the Lanchester Library, Coventry University.**

P3

This study demonstrated the challenges of working with real-world data generated from routine clinical care in India, particularly in resource-limited settings [75]. Firstly, data quality was ensured by inviting only centres providing the highest quality of care in India. In addition, best corrected VA (BCVA) was routinely collected in these hospitals, unlike most hospital data from LMICs which tend to report uncorrected VA. Furthermore, it was possible to track real world patient behaviour using this dataset; the number of patients that received timely treatment, those that were referred from screening programs and the outcomes of patients that received the best quality of care compared to those who didn't, over 10 years.

Secondly, both age-standardised (using direct standardization) and crude rates were provided which allows readers to assess the distortion caused by the age structure in the study sample relative to the 2001 India census data [76]. The formula for the direct standardised rate is as follows (**Equation 2**) [77] :

**Equation 2. Standardised rate using direct standardisation**

$$S_R = \frac{\sum_{i=1}^k w_i * R_i}{\sum_{i=1}^k w_i}$$

where  $w_i$  is the weight for stratum  $i$  derived from the standard population, and  $R_i$  is the stratum-specific observed rate in stratum  $i$ .

Thirdly, for robustness, both US and WHO definitions of VI and blindness were used to aid comparability with literature, as there are currently no internationally agreed criteria for VI and blindness.

## Results

### National prevalence of VI and blindness in PWD in India study

**P2** considered in total 7,910 participants aged 40 years and above with T2DM, of which 5,689 were known diabetes and 2,221 undiagnosed. The country-wide prevalence of VI in PwD was 21.1% (95% CI 15.7%-27.7%) and blindness 2.4% (95% CI 1.7%-3.4%) (**Table 4**), with no significant differences between known and undiagnosed diabetes. The proportion of ungradable images increased with worsening VA.

**Table 4. Estimated prevalence of VI based on US and WHO severity scale**

Criteria	Definition Snellen	logMAR equivalent	n	Estimate, % (95% CI) <sup>a</sup>
Normal vision	6/12 or better	≤0.3	6,571	78.9 (72.3-84.3)
<b>US criterion</b>				
VI	<6/12 & >6/60	>0.3 & <1.0	1,213	18.7 (13.7-24.9)
Blindness	6/60 or worse	≥1.0	126	2.4 (1.7-3.4)
Total (any VI)	<6/12	>0.3	1,339	21.1 (15.7-27.7)
<b>WHO criterion</b>				
Normal Vision	6/12 or better	≤0.3	6,571	78.9 (72.3-84.3)
Mild VI	<6/12-6/18	>0.3 & ≤0.5	756	11.6 (8.7-15.2)
Moderate VI	<6/18-6/60	>0.5 & ≤1.0	498	7.9 (5.0-12.0)
Severe VI/blindness	<6/60	>1.0	85	1.6 (1.1-2.5)
Total (any VI)	<6/12	>0.3	1,339	21.1 (15.7-27.7)

Abbreviations: VI -Vision impairment; US -United States; WHO -World Health Organisation.

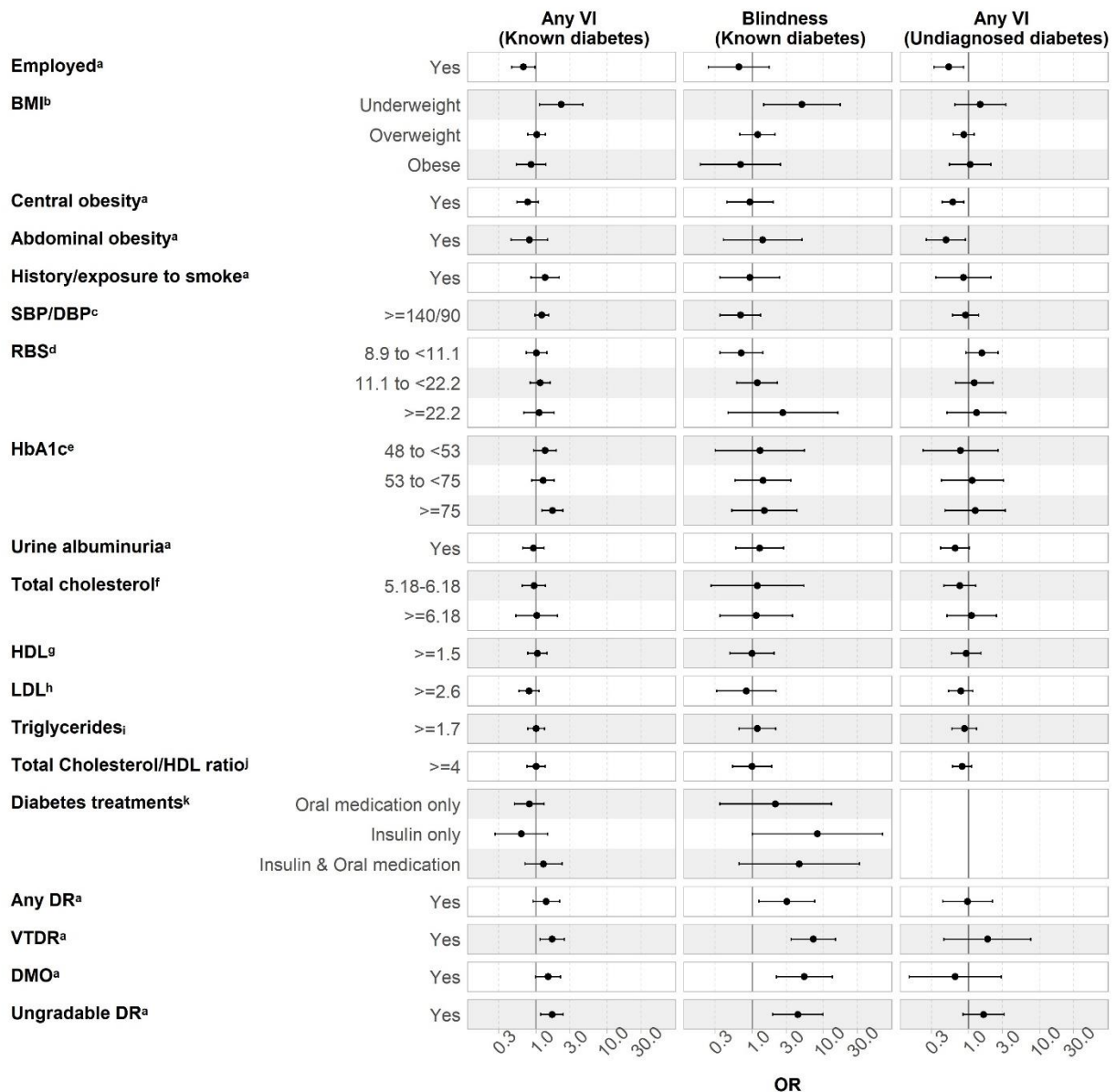
<sup>a</sup> Prevalence(95% CI) generated using the “logit” method of R survey package, which fits a logistic regression model, computing “Wald” intervals on the log-odds scale. Extracted from P2.

Moreover, this study showed that older age and lower educational attainment were common socio-demographic risk factors for any VI and blindness in both undiagnosed and known diabetes.



Among those with gradable DR, any DR, STDR and DMO had increased odds of blindness, while ungradable scans had greater odds of both any VI and blindness (Figure 5).

**Figure 5. Figure Odds Ratio plot showing risk factor burden for VI in known and undiagnosed diabetes from adjusted survey weighted logistic regression**



Variables age, gender, diabetes duration, rurality, ETL and education status were adjusted for in multivariable survey weighted logistic regression models. Any VI defined as VA≥0.4 logMAR, blindness defined as VA≥1.0 logMAR. Reference categories as follows; <sup>a</sup> No, <sup>b</sup> BMI 18.5 to < 25 kg/m<sup>2</sup> (normal), <sup>c</sup> <140/90 mmHg, <sup>d</sup> RBS<8.9 mmol/L, <sup>e</sup> HbA1c<6.5%, <sup>f</sup> Total cholesterol<5.18

mmol/L, <sup>g</sup> HDL<1.5 mmol/L, <sup>h</sup> LDL<2.6 mmol/L, <sup>i</sup> Triglycerides<1.7 mmol/L, <sup>j</sup> Total Cholesterol/HDL ratio <4, <sup>k</sup> Diet controlled. Abbreviations: VI – Vision Impairment, BMI- Body Mass Index, SBP/DBP- Systolic Blood Pressure/Diastolic Blood Pressure, RBS- Random Blood Sugar, HbA1c- glycated hemoglobin A1c, HDL- High Density Lipoprotein, LDL- Low density lipoprotein, DR- Diabetic Retinopathy, VTDR- Vision Threatening Diabetic Retinopathy, DMO- Diabetic Macular Oedema. Extracted from P2.

#### Prevalence and incidence of VI in PDR eyes following PRP in India study

For **P3**, the analysis sample included 516, 424 and 455 patients at baseline, 5 years, and 10 years respectively. 10-year crude incidence rates are provided in **Table 5**. The age-standardised incidence of VI at 10 years using the US criterion and WHO criterion was 14.2 (95% CI 7.1, 21.3) and 9.3 (95% CI 3.6, 14.9) respectively. The age standardised incidence of blindness at 10 years was 14.6 (95% CI 7.9, 21.4) and 14.6 (95% CI 7.7, 21.5) using the US and WHO criterion respectively. Eyes treated in the early stages of disease had better visual outcomes, supporting previous literature [78]. Moreover, patients referred from DR screening programs demonstrated a higher likelihood of being diagnosed in early stages (VA  $\geq$  6/12 Snellen). Another important point was that only a third of PDR eyes were referred from DR screening programs.

**Table 5. Age stratified ten-year crude incidence of Visual Impairment and Blindness based on best corrected visual acuity**

Age at baseline (years)	Incidence of Visual impairment			Incidence of Blindness		
	N	n	% (CI)	N	n	% (CI)
<b>United States Criterion</b>						
<40	34	4	11.8 (4.7, 29.5)	40	6	15 (7.2, 31.4)
40 – 49	91	17	18.7 (12.2, 28.7)	103	15	14.6 (7.2, 18.6)
50 – 59	164	28	17.1 (12.2, 23.9)	191	26	13.6 (9.5, 19.5)
60 – 69	81	16	19.8 (13.7, 30.6)	87	8	9.2 (4.8, 17.8)
70 +	3	1	33.3 (6.7, 165.1)	4	0	NA
			p = 0.39			p = 0.22
Crude Overall	373	66	17.7 (14.2, 22.0)	425	55	12.9 (9.5, 15.6)
<b>WHO Criterion</b>						

<40	38	4	10.5 (4.2, 26.6)	41	3	7.3 (2.5, 21.8)
40 – 49	98	13	13.3 (8.0, 22.0)	107	8	7.5 (3.8, 14.6)
50 – 59	181	21	11.6 (7.8, 17.3)	196	16	8.2 (5.1, 13.0)
60 – 69	85	7	8.2 (4.0, 16.7)	93	7	7.5 (3.7, 15.3)
70 +	3	0	NA	4	0	NA
			p = 0.40			p = 0.96
Crude Overall	405	45	11.1 (8.4, 14.6)	441	34	7.7 (5.6, 10.6)

*N= number at risk at baseline; n= incident cases; % (CI)= prevalence and 95 percent confidence interval.; p value calculated using Chi2-test of trend; Number of observations = 455. NA: No incident cases. Extracted and amended from P3 [58].*

## Originality and Contribution to the subject

Limited studies in India offer nationally representative prevalence data for VI and blindness in PwD and detailed estimates of their prevalence in varying DR severity. Studies using rich data sources that report the risk factor burden in VI and blindness are also scarce (**P2**). Moreover, there's a lack of follow-up data in LMICs, providing both short and long-term outcomes, on high-risk patients (**P3**). Key contributions to knowledge include:

- i) **P2** with national coverage provides the most detailed and latest estimates for prevalence of VI and blindness in varying grades of DR. The RAAB 2015-19 survey [66, 74] did not go into this level of detail and 2012 Sankara Nethralaya-Diabetic Retinopathy Epidemiology And Molecular genetic Study (SN-DREAMS) [79] estimates are outdated. These figures highlight the increasing need for diabetes and blindness prevention programs. Additionally, they aid policymakers in resource allocation and planning.
- ii) **P3's** findings were consistent with the literature, emphasizing baseline VA's predictive importance [80]. The study highlights several areas for growth in DR screening in secondary care in India, such as the emphasis on the variability in visual outcomes in people undergoing PRP, low DR referral rates and late presentation of disease. Moreover, solutions were proposed to reduce VI and blindness in PDR. Increased patient education and accredited social health activist (ASHA) workers have helped bridge the gap between the rural community and healthcare since 2005 [81], however

state-based healthcare disparities and personalised care must also be addressed [82, 83].

## Critical reflection

**P2** was by far the largest primary data collected by the ORNATE-India project, surveying a total of 42,146 participants with and without diabetes. My study focused on a subset of 7,910 participants with diabetes. Ungradable scans (approx. 22.5% in our study) tend to be higher in nonmydriatic screening [84], with a prior study reporting cataract in 40% of these images [85]. In India cataracts are the leading cause of VI [66]. Going forward, better, and more affordable handheld devices should be developed for house-to-house survey use. The study cohort focused on individuals aged 40 years and above to ensure comparability with regional surveys. Future research should consider including 20–40-year age group, a significant proportion of India's working age population. Furthermore, although not all states were sampled, the inclusion of sites from New Delhi may compensate for the absence of samples from Uttar Pradesh (the largest state), as a significant portion of immigrants in New Delhi originate from these states.

The Coronavirus Disease-19 (COVID-19) pandemic halted the recruitment for the SMART-India study in March 2020. Despite these unexpected setbacks, swift communication and prompt image analysis enabled successful project completion. Overall, the SMART-India study was a major contributor to the success of the Ornate India project, and the findings from Kerala resulted in a policy change in the state whereby diabetic retinal screening is now mandatory for those on the diabetes register [86].

While **P3** focuses on PDR, the study provided both long-term and short-term data on vision. I produced the revised statistical analysis plan in this study, including the standardisation of the rates based on the age structure of the 2001 Indian census population, as well as additional multivariable analysis for risk factor evaluation. Despite its retrospective nature, BCVA data quality was good, minimising underestimation bias [87, 88]. Transitioning to Electronic Health Records (EHRs) in all Indian states, is recommended for improved data quality, standardization, and to facilitate clinical research. A final critique would be the sample was limited to specialist retinal services in India; future research should consider a comprehensive nationwide study.

In summary, both studies offer a thorough evaluation of the impact of DR and diabetes on vision. **P2** provides national estimates of VI and blindness prevalence in PwD in India and revealed socioeconomic disparities and increased prevalence in DR. Meanwhile, **P3** highlights the concerning prevalence and incidence of VI and blindness in eyes with PDR despite PRP,

emphasizing the urgency of diabetes prevention initiatives and early intervention programs. This highlights the urgency of our DR risk stratification research (**chapters 4-6**).

## Chapter 3. Ethnic differences in DR

### Context and Objective

As ethnic minorities exhibit a higher DR prevalence compared to their white counterparts, targeting these at-risk individuals should underpin DR prevention and treatment. Yet, there is a scarcity of data regarding disparities in the incidence of DR and STDR within ethnic minority groups. Recent studies conducted on nationally representative clinical practice research datalink (CPRD) data identified over a third of patients with missing ethnicity records. East London GP's have seen an annual rise in ethnicity recording to levels of 80% in 2010 [89], where ethnic minorities make up around 50% of the population in this region. During the same period, recording of those on the diabetes register and other chronic disease registers reached 98% of the East London registered population, making it an ideal cohort to study ethnic minority groups in detail.

**Publication 4 (P4)** explores the health inequalities seen in UK ethnic minority groups and determines the risk factors for DR and STDR in PwD from the GP data of East London, independent of ethnicity and/or race (**Table 6**).

**Table 6. Chapter 3 Publication 4 with citations and mentions, updated 16/03/2024.**

P4: Ethnic differences in DR study

Ethnic Disparities in the Development of Sight-Threatening Diabetic Retinopathy in a UK Multi-Ethnic Population with Diabetes: An Observational Cohort Study

Nugawela MD, **Gurudas S**, Prevost AT, Mathur R, Robson J, Hanif W, Majeed A, Sivaprasad S

2021, JPM, Impact factor: 3.508



### Methodological commentary and Critical powers

The Strengthening the Reporting of Observational Studies in Epidemiology or “STROBE” guidelines was followed to ensure methodological rigour [90]. In the UK, since PwD are monitored frequently, real-world data sources tend to be rich containing information on biochemical parameters, physical examinations, co-morbidities, prescription medication and lifestyle. Study design considerations to minimize potential for bias include accurate cohort selection by ensuring relevant read-codes were extracted, limiting miss-classification bias on diagnoses, missing measurements, definition of censoring and categorisation of variables. Data quality assessments were done manually (concordance, correctness, plausibility, completeness and currency) to

improve the credibility of the dataset [91]. The following criteria were considered in all publications that used this dataset (**Table 7**).

**Table 7. EHR data quality checks**

<b>Assessment criteria</b>	<b>Definition</b>	<b>Case of assessment criteria and exclusions to address the case</b>
Concordance	Agreement between data elements	Precision of date of T2DM diagnosis given; coincides with earliest T2DM date given; patients are not taking antidiabetic medication prior to diagnosis date; patients are not prescribed 2 antidiabetic medications or insulin on date of diagnosis of T2DM
Correctness	A value is true	Misclassification of T2DM as T1DM; by excluding patients who are prescribed 2 antidiabetic medications or insulin on date of diagnosis of T2DM. Precision of time to event; by setting robust study start (latest of date of 18 <sup>th</sup> birthday, 12 months after registration, or January 2007) and follow-up end date (earliest of death, de-registration, latest data collection, or January 2017).
Plausibility	A value is plausible based on external information	Each record for each covariate was examined in relation to possible values, e.g., systolic blood pressure to fall between values 70 and 240
Completeness	A truth about a patient is present	Missing data in covariates; excluded patients that did not have a record within +/-6 months to the study baseline date, ethnicity was recorded in 98% individuals (~1.8% missing an ethnicity record which were excluded)
Currency	A value is representative of the clinically relevant time	Closest record for the covariates to the study baseline date (+/- 6 months)

*Abbreviations; T2DM- type 2 diabetes mellitus, T1DM-Type 1 diabetes mellitus, EHR- Electronic health record. Assessment criteria based on recent guidelines[91] and amended to suit P4.*

Cox proportional hazards model (**Equation 3**) was used to model 10-year incidence of DR and STDR, a robust method widely used in DR research due to its versatility and intuitive relative-risk interpretation [92, 93].

**Equation 3. Specification of the Cox proportional hazards model for the hazard rate**

$$h(t; x) = h_0(t) \exp(x(t)\beta)$$

$t$  – time from start date

$h(t)$  – hazard function

$h_0(t)$  – Baseline hazard function

$\beta$  is a vector of unknown regression parameters

The hazard function  $h(t; x)$  describes the instantaneous risk of experiencing an event within an infinitesimal interval of time, given the event has not occurred. It relies on proportional hazards, where covariate hazard ratios remain constant over time. The model coefficients closely resemble the correctly specified parametric model without need for the specification of a functional form for the baseline hazard,  $h_0(t)$ . Moreover, the relative-risk interpretation of the parameters from a Cox model is intuitive and improves interpretability of the study results [94]. Multivariable analysis adjusted for demographic, socio-economic status (defined by Townsend deprivation score), diabetes duration, laboratory data, blood pressure, anthropometric measures, co-morbidities, and medication history to reduce potential for confounding and draw robust causal conclusions. Adjusting for Townsend deprivation index, a measure of area deprivation incorporating information on unemployment, overcrowding, home ownership, and car or van ownership, controls confounding effects due to area-based socioeconomic differences which may be confounded with ethnic group. Finally, the study comprised both newly diagnosed and prevalent (those who entered the study after diagnosis of diabetes) PwD, selecting the first diabetes appointment between 2007 and 2017 as the baseline date. Approximately 50% of participants had their baseline set at the onset of T2DM. Many of the studied covariates may not apply to both groups, and therefore cohorts were analysed separately.

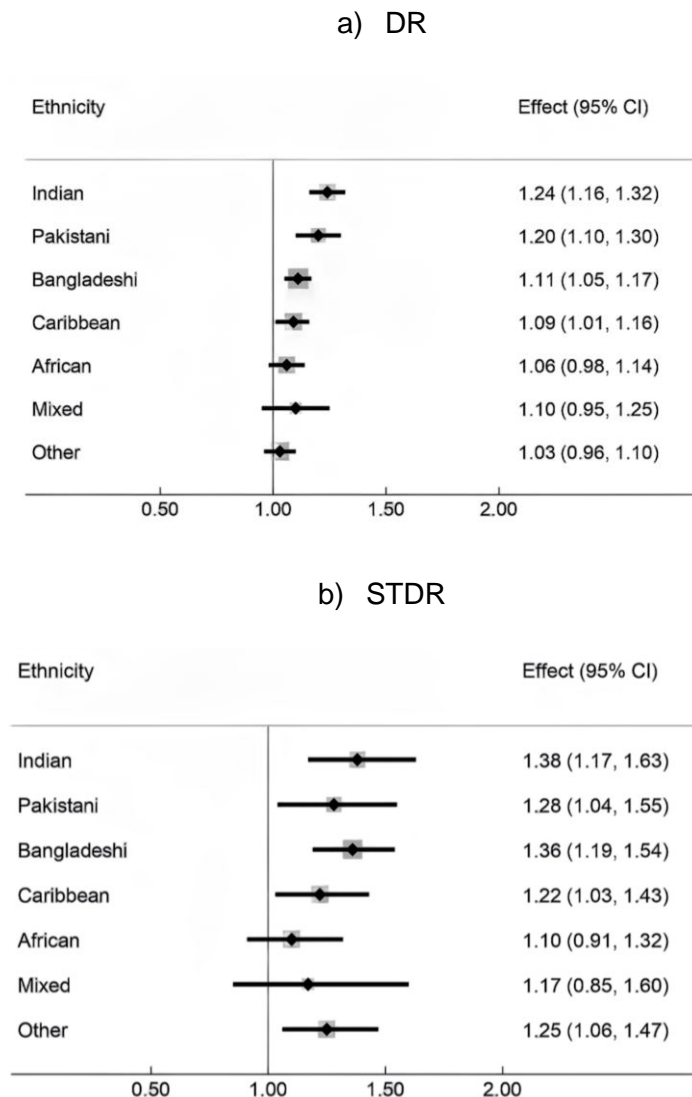
## Results

EHRs of the eligible 52,216 PwD from 134 GPs in East London between 2007–2017 were analysed. UK ethnic minorities (Indian, Pakistani, Bangladeshi and Caribbean) were found to have increased risk of DR and STDR compared to their white counterparts in both newly diagnosed and prevalent T2DM. Indians had the highest risk of any DR (adjusted HR 1.24 [95%



CI 1.16-1.32]) and STDR (adjusted HR 1.38 [95% CI 1.17-1.63]) in prevalent T2DM in multivariable analysis (**Figure 6**). Among South Asians, and Black individuals, this risk varied by their ethnic subgroup. Moreover, increased diabetes duration, male sex, uncontrolled blood glucose, hypertension, kidney impairment, use of insulin or 2 anti-diabetic drugs, were at increased risk of incident DR and STDR. Use of statins was additionally associated with incident DR. In univariate analysis, despite increased risk of incident DR and STDR, South Asians were generally younger, affluent and had lower blood pressure compared to all other ethnicities.

**Figure 6. Illustrating the risk of DR and by ethnicity in prevalent T2DM at baseline**



*Abbreviations: DR-Diabetic retinopathy; STDR- Sight threatening diabetic retinopathy; CI- Confidence interval. Effect refers to Hazard Ratios, with white ethnicity as the reference category and time censored at 10-years. Extracted from P4 [59].*

## **Originality and Contribution to the subject**

Existing studies on ethnic disparities in DR and STDR, such as CPRD data (2004-2014) in DR had high proportion of missing data for ethnicity [95]. Our study provides contemporary estimates for the incidence of DR and STDR in both newly diagnosed and prevalent T2DM, and its relationship with key risk factors.

This work makes three notable contributions. Firstly, it quantifies DR and STDR incidence in a diverse UK cohort from 2007 to 2017, well-powered to study minor ethnic groups. Secondly, it reaffirms previous findings, that South Asians and Black individuals have a higher risk of any DR and STDR compared to their white counterparts after adjusting for known risk factors [96]. The analysis highlighted differences in newly diagnosed and prevalent T2DM. Africans were 36% more likely to have STDR compared to individuals who identified as white in newly diagnosed T2DM, while among those with prevalent T2DM, Indians exhibit a 24% increased risk of DR and a 38% increased risk of STDR compared to white individuals. This underscores the importance of recording self-reported ethnicity, now a contractual requirement in the 2020 GP contract and for hospitals [97, 98], to ensure better national data representation and data quality. Furthermore, these associations remain independent of the key DR risk factors and Townsend deprivation score, suggesting potential ethnic influence on DR pathophysiology, aligning with emerging genetic research on DR [99]. Adjusting for Townsend deprivation scores demonstrates that our findings are not driven by differences in socioeconomic status. Moreover, despite South Asians' younger age, affluence and lower blood pressure compared to other ethnicities, a higher incidence of DR and STDR was observed in multivariable analysis. Thirdly, this study establishes a set of independent risk factors for incident DR and STDR in addition to ethnicity, emphasising the need for multifaceted T2DM prevention strategies. Policymakers can use these findings to target high-risk groups for improved care strategies.

## **Critical reflection**

The study rigorously explores and builds on the evidence from the global literature review (P1), addressing challenges in DR research, emphasising the need for LMIC and minor ethnic group investigations. It includes an extensive set of candidate risk factors, identified from prior literature, enabling the adjustment for chance imbalances. In addition, as the NHS is free to all and provides standardised clinical care, bias in our study due to care inequalities would be minimal. Although Townsend Index was incorporated into the multivariable analysis to address care inequalities due to area deprivation, it is recognised that this measure does not fully grasp the intricacies of

socioeconomic status. It does not consider the combined effects of affluence and deprivation within a community. Individual-level socioeconomic indicators can provide a more accurate representation of health inequalities, supplementing area-level indicators which may underestimate individual poverty levels [100-102]. Moreover, previous research has shown that poorer individuals living in deprived neighborhoods often suffer the most severe health consequences [103], emphasizing the need to consider both individual-level indicators and area deprivation to capture varying health inequalities, a key limitation of our analysis. Similar to previous research, our findings support the need for ethnicity-based DR screening in the NHS [104], warranting randomised control trials to assess the utility and cost-effectiveness of such screening programs in the NHS. While the semi-parametric (Cox) approach has been commended for its flexibility and interpretability, it can be more computationally demanding, especially on larger datasets, than parametric models that have an assumed hazard function and thus have a less complex estimation process. Parametric models which parametrise the baseline hazard, are the more efficient approach when the chosen distribution aligns well with the data [105, 106]. Potential biases that can arise from model misspecification include incorrect parameter estimates (lack of consistency), incorrect standard errors of the estimated regression coefficients (lack of efficiency), biased predictions, loss of flexibility and overfitting [107]. An interval-censored approach may more accurately estimate incidence of DR in routine healthcare data than traditional Cox models due to the periodic follow-up, as the event is assumed to have occurred in an interval rather than observed exactly[108]. At the time of conducting the analysis for this study, the interval-censored cox model was nascent in software and became available in Stata 17[109], therefore I did not consider this model over the Cox proportional hazards models. The interval-censoring approach was explored in **P7** which uses the same participant cohort and found comparable incidence rates for STDR with the Cox model, possibly due to low rate of interval censored outcomes.

In summary, this study uncovers the ethnic disparities and factors associated with DR incidence including increased diabetes duration, male sex, uncontrolled blood glucose, hypertension, kidney impairment, use of insulin or 2 anti-diabetic drugs. These findings can be used to inform tailored care strategies and risk-based screening.

## Chapter 4. Diagnostics tests for DR

### Context and Objective

Many circulating biomarkers are shown to be expressed in DR [110-114]. The prognostic utility of the extent of kidney impairment (eGFR), hypertension and hyperglycemia were already demonstrated in **P4**. However, their relative contributions compared to known risk factors hyperglycaemia, hypertension, and hyperlipidemia [115, 116] remain limited.

Studies by the Kim group [52, 117] explored potential biomarkers using serum samples for different DR severities, however larger studies are needed to verify their findings. Our studies aimed to identify one or more biomarkers for integration into an affordable biosensor, enabling real-time at-home risk factor monitoring. Affordability of biosensors is crucial for LMICs; thus, our methodological decisions prioritise accuracy while minimising the number of markers needed.

Publication 5 (**P5**) assesses the relative contributions of many circulating biomarkers and environmental factors for DR in NHANES, offering an unbiased model-agnostic approach to identify cost-effective biomarkers for DR detection. Publication 6 (**P6**) examines whether any one or combination of 13 circulating biomarkers identified in the literature, could be used to distinguish STDR from no DR in the UK and India (**Table 8**).

**Table 8. Chapter 4 publications 5 and 6 with citations and mentions, updated 16/03/2024.**

P5: Environment-wide association study (EWAS) for DR

Diabetic Retinopathy Environment-Wide Association Study (EWAS) in NHANES 2005-2008

Blighe, K., **Gurudas, S.**, Lee, Y., & Sivaprasad, S  
2020, JCM, Impact factor: 4.964



P6: Circulating biomarkers in STDR study

Multicenter Evaluation of Diagnostic Circulating Biomarkers to Detect Sight-Threatening Diabetic Retinopathy

**Gurudas S\***, **Frudd K\***, Maheshwari J, Revathy YR, Sivaprasad S, Ramanathan S, Pooleeswaran V, Prevost AT, Karatsai E, Halim S, Chandra S, Nderitu P, Conroy D, Krishnakumar S, Parameswaran S, Dharmalingam K, Ramasamy K, Raman R, Jones C, Eleftheriadis H, Greenwood J, Turowski P  
2022, JAMA Oph, Impact factor: 8.3



## Methodological commentary and Critical powers

**P5** conducted an EWAS [118] with over 400 variables on two NHANES waves (2005-06 and 2007-08), employing a data-driven approach to examine their relative contributions and relationships. In contrast, **P6**, based on an exhaustive literature review, focused solely on biomarkers identified in the review [30]. Both were cross-sectional studies using logistic regression for modelling DR. In **P5**, survey weighted logistic regression was applied (**Equation 1**), using the prescribed survey weights from the demographic documentation of the NHANES survey [119]. The design features of NHANES are crucial to ensuring sample representativeness. NHANES uses a four-stage sampling design. This includes stratification; dividing the US into strata based on census regions and geographic information. Primary sampling units (PSU's) which are the US counties were randomly selected within each strata in the first stage, with census blocks selected proportional to size within counties in the second stage, households randomly selected within census blocks oversampling certain groups (based on age, race/ethnicity and income) in the third stage and then individuals randomly selected within census blocks in the fourth stage.. Probability sampling assigns known (unequal) probability of selection to individuals. Oversampling ensures that subpopulations have adequate sample size, with special effort made in NHANES to oversample racial and ethnic minority groups and other special populations. Weighting was applied in the analysis to address oversampling and nonresponse.

Due to the large number of variables, systematic data cleaning was performed, involving the removal of variables with over 90% missing data and the exclusion of categorical variables incompatible with the Principle Component Analysis (PCA) approach and detrimental to Random Forest's (RF) splitting points. Previous research utilizing the NHANES dataset addressed this issue by excluding variables with >90% missing data [120]. While this approach may result in the inclusion of variables with a substantial proportion of missing data, it ensured a satisfactory sample size for the adjusted logistic regression analysis while preserving a considerable number of variables for the EWAS analysis. Multiple imputation (MI) techniques were not employed due to the extensive number of variables (>400), as MI is recognised to pose challenges in high-dimensional datasets. It has been noted that likelihood estimates may encounter convergence issues as the number of variables approaches the sample size [121]. Laboratory variables underwent z-score transformation for uniform scaling and outlier mitigation, confounders (age, ethnicity, and diabetes duration) were adjusted for in logistic regression analysis pre-identified from literature, sampling weights were applied [70] to preserve national representativeness and false discovery rate (FDR) was controlled for using the Benjamin-Hochberg procedure. To reduce

model complexity and likelihood of multicollinearity, elastic net regularization (penalised regression with L1 and L2 penalties) was used for variable selection [122]. For both penalised regression and RF, cross-validation was used to minimise the risk of overfitting bias. Receiver operating characteristic (ROC) curves were used to measure marker discriminative abilities after grouping them by their biological pathway. Moreover, the study results were replicated in an independent NHANES cohort.

Analysing existing datasets like NHANES offers analytical flexibility due to its large sample size and the ability to assess numerous variables at low cost, but risks increased type I errors. In contrast, **P6** utilised primary data from outpatient ophthalmology clinics in the UK and India, following a consistent study design, selecting candidate biomarkers based on prior evidence, unavailable in existing data sources. This strengthened the scientific merit of our findings, in contrast to data driven methods used in **P5** that carry an increased risk of false positives. However, **P5's** methods provide a valuable foundation for generating new hypotheses. **P6** combined insights from a comprehensive literature review [30] that pinpointed the biomarkers under study, and primary data from two independent cohorts, which allowed the exploration of intercountry variations in biomarker profiles, while maintaining consistency in laboratory methods and data collection practices. To address potential bias from over-sampling STDR groups (NPDR with DMO and PDR), weighted logistic regression and weighted ROCs were used to mitigate differences in the sample and population prevalence. Probability weights  $w_i$  for each disease group, were calculated by dividing the population proportion ( $\pi_i$ ) (based on a UK sample [123] and SMART-India data [69]), by the sample proportion ( $\bar{y}_i$ ). The likelihood contributions, like **P5** were weighted according to **Equation 1**. Forward stepwise selection (entry criterion:  $\alpha = 0.1$ ), with age, disease duration, ethnicity (in the UK), and  $HbA_{1c}$  was applied on significant variables from the adjusted analysis. Fractional polynomial terms were considered for non-linearity. Patients with NPDR without DMO were not recruited due to time and cost constraints, thus our results pertain to a population excluding this group. In both studies, adherence to the STROBE guidelines was ensured. Additionally, the STARD (Standards for Reporting Diagnostic accuracy studies) guidelines were met. These guidelines offer a comprehensive set of criteria aimed at enhancing the transparency of reporting in diagnostic accuracy studies (**Appendix 2**).

## Results

### P5: Environment-wide association study (EWAS) for DR in NHANES

A total of 1025 eligible participants with diabetes in National Health and Nutrition Examination Survey (NHANES) 2007–2008 wave and 637 participants from 2005-2006 wave were included. Over 400 laboratory parameters were assessed and compared with the established risk factors for DR. Statistically significant risk factors from adjusted logistic regression were reported in **Table 9**. HbA1c was the strongest ranking circulating biomarker in several independent analyses (PCA, penalised regression and RF).

**Table 9. Statistically significant laboratory variables following FDR correction, associated with DR in PwD**

Variable	Age*		Ethnicity*		Diabetes duration*	
	OR (95% CI)	p-value**	OR (95% CI)	p-value**	OR (95% CI)	p-value**
Glycohemoglobin (%)	2.34 (1.87-2.92)	0.001	2.28 (1.82-2.87)	0.004	2.05 (1.55-2.73)	0.03
Glucose, serum (mmol/L)	1.77 (1.45-2.15)	0.01	1.71 (1.41-2.09)	0.03	1.43 (1.1-1.86)	0.05
Osmolality (mmol/Kg)	1.57 (1.3-1.9)	0.01	1.63 (1.34-1.99)	0.07	1.48 (1.12-1.94)	0.04
Albumin, urine (mg/L)	1.53 (1.24-1.88)	0.01	1.53 (1.25-1.87)	0.14	1.28 (0.98-1.68)	0.16
Hemoglobin (g/dL)	0.74 (0.63-0.88)	0.02	0.75 (0.62-0.9)	0.26	1 (0.73-1.37)	0.99
Fasting Glucose (mmol/L)	1.66 (1.3-2.13)	0.01	1.59 (1.25-2.02)	0.23	1.24 (0.93-1.66)	0.27
NNAL , urine (ng/mL)	0.82 (0.71-0.94)	0.06	0.75 (0.65-0.87)	0.23	0.78 (0.52-1.19)	0.41
Iodine, urine (ug/L)	0.81 (0.73-0.9)	0.02	0.86 (0.77-0.96)	0.26	0.75 (0.63-0.89)	0.03
Cobalt, urine (ug/L)	0.6 (0.45-0.82)	0.03	0.6 (0.44-0.81)	0.26	0.59 (0.45-0.78)	0.03
Hematocrit (%)	0.76 (0.62-0.92)	0.06	0.75 (0.62-0.92)	0.26	0.97 (0.7-1.35)	0.93
Blood urea nitrogen (mmol/L)	1.31 (1.01-1.71)	0.16	1.43 (1.16-1.75)	0.26	1.17 (0.87-1.57)	0.48
Albumin (g/L)	0.81 (0.69-0.94)	0.07	0.83 (0.7-0.99)	0.32	0.92 (0.74-1.14)	0.59
Urinary Triclosan (ng/mL)	0.67 (0.5-0.89)	0.06	0.66 (0.49-0.89)	0.26	0.55 (0.36-0.83)	0.04
Mean cell hemoglobin (pg)	0.77 (0.65-0.91)	0.04	0.84 (0.7-1.02)	0.49	0.99 (0.76-1.28)	0.95



Lead, urine (ug/L)	0.67 (0.49-0.91)	0.08	0.66 (0.49-0.9)	0.26	0.65 (0.43-0.99)	0.13
Creatinine, urine (umol/L)	0.83 (0.68-1)	0.17	0.77 (0.64-0.92)	0.26	0.76 (0.64-0.9)	0.03
Alanine aminotransferase (ALT) (U/L)	0.82 (0.66-1.02)	0.20	0.81 (0.64-1.01)	0.43	0.92 (0.68-1.23)	0.70
Creatinine (µmol/L)	1.19 (0.96-1.48)	0.25	1.23 (0.99-1.53)	0.44	1.14 (0.89-1.45)	0.47
Red blood cell count (million cells/uL)	0.88 (0.74-1.04)	0.27	0.83 (0.7-0.98)	0.32	1.01 (0.74-1.37)	0.98
Mean cell volume (fL)	0.78 (0.65-0.93)	0.06	0.86 (0.71-1.06)	0.68	0.95 (0.72-1.26)	0.84
Platelet count (1000 cells/uL)	0.83 (0.68-1.02)	0.20	0.79 (0.67-0.94)	0.26	0.75 (0.57-0.99)	0.13
Mean platelet volume (fL)	1.3 (1.06-1.59)	0.08	1.25 (1.02-1.53)	0.32	1.09 (0.83-1.43)	0.66
Cotinine (ng/mL)	0.91 (0.79-1.04)	0.34	0.84 (0.74-0.95)	0.26	0.94 (0.66-1.34)	0.84
Insulin (pmol/L)	0.74 (0.55-1)	0.17	0.75 (0.56-1)	0.38	0.8 (0.51-1.26)	0.51
Blood cadmium (nmol/L)	0.78 (0.64-0.95)	0.09	0.79 (0.64-0.98)	0.32	0.76 (0.5-1.16)	0.35
Urinary perchlorate (ng/mL)	0.78 (0.63-0.98)	0.13	0.78 (0.64-0.96)	0.27	0.73 (0.53-1.01)	0.15
Urinary nitrate (ng/mL)	0.79 (0.63-1)	0.16	0.76 (0.6-0.96)	0.27	0.71 (0.55-0.91)	0.05
Cesium, urine (ug/L)	0.72 (0.55-0.96)	0.11	0.72 (0.54-0.95)	0.27	0.71 (0.47-1.06)	0.20
Thallium, urine (ug/L)	0.68 (0.48-0.96)	0.13	0.66 (0.47-0.92)	0.26	0.62 (0.4-0.94)	0.09
25OHD2+25OHD3 (nmol/L)	0.78 (0.65-0.94)	0.08	0.83 (0.66-1.04)	0.57	0.82 (0.61-1.1)	0.34
Blood Toluene (ng/mL)	0.82 (0.69-0.97)	0.10	0.8 (0.65-0.98)	0.32	0.74 (0.45-1.21)	0.39

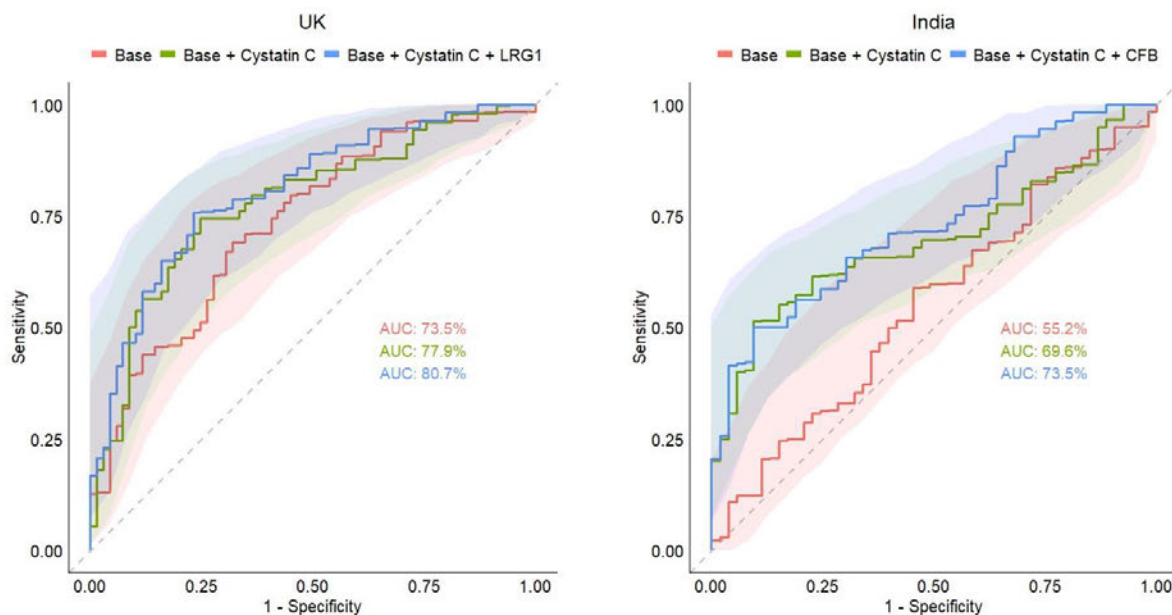
C-reactive protein(mg/dL)	0.83 (0.69-1.01)	0.18	0.78 (0.66-0.92)	0.26	0.77 (0.6-0.98)	0.11
Barium, urine (ug/L)	0.69 (0.48-0.99)	0.15	0.68 (0.47-1)	0.39	0.64 (0.47-0.87)	0.04
Urinary 4-tert-octylphenol (ng/mL)	0.72 (0.46-1.12)	0.30	0.66 (0.43-1.01)	0.42	0.44 (0.23-0.87)	0.08
Dimethyldithiophosphate (ug/L)	0.71 (0.49-1.01)	0.18	0.81 (0.59-1.12)	0.74	0.51 (0.29-0.91)	0.09

Abbreviations: OR- Odds Ratio; CI- Confidence Interval. \*Variable adjusted in models. \*\*False discovery rate (FDR) adjusted p-value.  
 Extracted and modified from P5 [60].

## P6: Circulating biomarkers in STDR, UK- and India study

The diagnostic abilities of biomarkers were compared and found that no other biomarker outperformed Cystatin-C in both UK (N=215) and India participants (N=208). ROC analysis confirmed that Cystatin-C (with age, disease duration, ethnicity (in the UK) and HbA1c) discriminated well between STDR and no DR in both countries (Area Under the Curve (AUC) 0.779 [95% CI 0.700-0.857] in the UK and 0.696 [95% CI 0.602-0.791] in India, **Figure 7**). While LRG1 and CFB improved model AUC in UK and India samples respectively, the improvement was only marginal relative to inclusion of Cystatin-C.

**Figure 7. Diagnostic performance of biomarkers for STDR selected from a forward stepwise routine with AUC-ROC curves and 95% CIs for UK and India**



*Abbreviations: UK- United Kingdom, LRG1- Leucine-rich  $\alpha$ -2 glycoprotein 1, CFB- Complement Factor B, AUC- Area Under the Curve, ROC-Receiver operating characteristic curves. Base model includes age, diabetes duration, HbA1c and ethnicity (in the UK models). In the UK, n=215 patients with 146 sight threatening diabetic retinopathy events. In India, n=208 patients with 155 sight threatening diabetic retinopathy events. Modified from P6 [61].*

## **Originality and Contribution to the subject**

**P5** demonstrates the first comprehensive study investigating over 400 laboratory markers, for their diagnostic ability in detecting DR. Notably, it reaffirmed evidence on HbA1c, traditionally the gold standard for assessing glycemic control and risk of diabetes complications, backed by an

independent NHANES replication (2005-06). Inflammatory markers occupy the hierarchy in the RF after hyperglycemia. Hypertension (elevated systolic blood pressure) also ranked highly in both penalised regression and RF algorithms, reinforcing prior literature. This study motivates continued biomarker research in DR as it revealed mechanisms that may be of interest (diabetes; immune status; renal function; haematocrit; Toxins/Metals; Sterols; Liver Function; Blood pressure).

**P6** was the largest study to have collected primary data on DR biomarkers, in 2 independent cohorts representing different health systems. Unlike previous studies, **P6** emphasised reproducibility and replicability of findings. It has significant implications to the field, as it paves the way for the development of a practical biosensor. An ongoing project in India achieved over 85% prediction accuracy using an electrochemical Cystatin-C sensor, validating its utility in both the UK and India [124]. In our study, the validity of Cystatin-C was similar in the UK and India, increasing the scientific merit of our findings. Finally, both **P6** and **P5** despite methodological differences, identified renal markers as valuable for DR classification. While Cystatin-C was not studied in **P5**, serum creatinine, closely related to and correlated with Cystatin-C, occupied the hierarchy in the RF model after HbA1c. This relationship was also alluded to in **P8**, showing that presence of STDR is associated with 5-year incidence of DKD.

## Critical reflection

**P5** was my first experience applying EWAS methodology and demonstrates my capabilities in teasing out associations on high-dimensional datasets (“wide data”). By drawing on results from **P4** and **P1** which identified key risk factors for DR (ethnicity, diabetes duration and HbA1c), bias due to confounding was minimised. In this study, I found that machine learning algorithms, used to extract patterns in rich, high-dimensional data sources, produced more fruitful analysis when combined with conventional statistical approaches to make inferences [125, 126]. A limitation of such data-driven studies is that the study was less streamlined, due to large number of variables. However, adjusting for the FDR, and using logistic regression to quantify relative risk combined with elastic net regularization for variable selection helped mitigate such biases. Our longitudinal studies (**P4**, **P7** and **P8**) were better equipped to make assertions on both associations at presentation and causality unlike the present studies which are cross-sectional in design. While EWAS methods are relatively new, they parallel the design, methodology and replicability standards of Genome-Wide Association Study (GWAS) methods which have been long established. Still, there is poor consensus on the quality control of environmental factors for statistical analysis [116, 127] and several previously identified biomarkers were unavailable in the

NHANES dataset, so further work would be needed to rule out the importance or efficiency of these biomarkers over and above HbA1c.

**P6** showed that Cystatin-C levels may be used to prioritise screening to identify people with a high likelihood of STDR. A limitation mentioned in a recent study citing our work, noted the absence of investigations into extracellular vesicles from urine and retinal tissue [128]. Our study, however, did consider prominent blood-based markers from serum samples [30]. One challenge was the loss of initial blood samples due to freezer failure, leading to time constraints. Nevertheless, the study promptly regained approval for data collection, ensuring successful completion. Following publication, the principal investigator and corresponding author S.S. held an in-depth discussion of our findings on Author Interviews Podcast from JAMA Ophthalmology, a podcast exploring the latest clinical research, views and opinions featured in the journal [129]. Future investigations should consider novel biomarkers not investigated in this study in addition to Cystatin-C. Both **P5** and **P6** performed ROC analysis with ROC's presented allowing readers to assess the sensitivity and specificity of the models corresponding to a threshold probability, however a critique of the AUC is that it summarises the entire ROC curve including regions that may not have any practical value. Partial AUC which summarises a section of the models ROC over a pre-specified sensitivity or specificity interval could alternatively be assessed. This can be interpreted as the average sensitivity (or specificity) in the pre-specified specificity (or sensitivity) interval.

In summary, findings from **P5** corroborate the evidence on HbA1c in a data-driven study investigating over 400+ markers in DR. While **P6** showcases the strengths of Cystatin-C not investigated in **P5**, and its utility as a screening tool for STDR.

# Chapter 5. Prognostic Modelling in Sight Threatening Diabetic Retinopathy

## Context and Objective

Preventing DR is largely driven by control of blood glucose and blood pressure, however these risk factors do not alone explain risk of STDR [130]. At the time of publication, 18 prognostic models were identified that required a previous record of retinopathy status, HbA1c test or other clinical or laboratory parameters, reducing its usability in LMICs. While **P5** and **P6** propose and validate biomarkers in PwD to classify STDR, these need to be developed into point-of-care tests to support self-testing and provide risk assessments for diagnosing existing STDR. The aim of **P7** was to develop and validate risk models that can be used to predict 3-year risk of STDR in any resource setting (**Table 10**). **P7** develops risk models using East London's GP-registered T2DM population (2007-2017), the same participants used in modelling for **P4**.

**Table 10. Chapter 5 Publication 7 citations and mentions, updated 16/03/2024.**

P7: STDR prediction modelling study

Development and validation of predictive risk models for sight threatening diabetic retinopathy in patients with type 2 diabetes to be applied as triage tools in resource limited settings

**Nugawela MD, Gurudas S,** Prevost AT, Mathur R, Robson J, Sathish T, Rafferty JM, Rajalakshmi R, Anjana RM, Jebarani S, Mohan V, Owens DR, Sivaprasad S  
2022, Lancet EclinicalMedicine, Impact factor: 17.033



## Methodological commentary and Critical powers

This study focused on statistical modelling for prediction, differing from previous publications that aimed at explaining causation and structure (aetiological inference). The Transparent Reporting of multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement for reporting prediction modelling studies was followed [131] with additional detailed guidance from “The PROgnosis RESearch Strategy” (PROGRESS) series specific to prediction modelling research [132].

Several predictors were identified in **P4**, however in this study the practicality of routinely collected data were considered in facilitating community DR screening. A comprehensive search from January 1980 until June 2021 on PubMed and Google Scholar revealed that prediction models need to be low-cost, low-maintenance, broadly applicable, accurate and address clinical need in

LMICs. A 2019 systematic review of STDR prognostic modelling studies identified predictors like HbA1c, diabetes duration, retinopathy presence, gender, age and systolic blood pressure appeared frequently in prognostic modelling studies [44]. Models were developed using the datasets used in **P4** and **P7** and validated with data from Wales (SAIL) and Chennai, India (MDRF). Discrimination was assessed using Harell's c-statistic, and calibration assessed using the beta coefficient of the linear predictor (calibration slope) and observed risks to expected risks ratio (O/E). Backward elimination (eliminating predictor variables least significant) was performed to select the final parsimonious models. Variables in Model 1 included age, gender, T2DM duration, antidiabetic medication use, HbA1c at baseline ( $\pm$  6 months), and presence of background retinopathy. Invasive tests HbA1c and background retinopathy were considered for elimination from the model. Model 2 excluded background retinopathy as the variable with least contribution to the c-statistic and Model 3 excluded both HbA1c and background retinopathy from Model 1.

The study design accounted for the timing misalignment between retinal screening and DR diagnosis by incorporating a 6-month delay between events in the development cohort. When no recorded DR event occurred within 6 months of a screening event, it was considered as evidence for absence of disease upon screening. The final DR screening date was taken as the last follow-up date (right censoring) to ensure accurate participant censoring. These participants that did not observe an event during follow-up constituted over 95% of the sample and therefore would have a large impact on the model coefficients if mis-specified.

Furthermore, final models initially estimated using Cox regression, were re-estimated, accounting for interval-censoring present in routine screening data, using interval-censored cox models to avoid underestimation of time-to-event. The Cox proportional hazards model (**Equation 3**), estimates model parameters by maximizing the partial likelihood function. In the interval-censored approach, a novel expectation-maximisation (EM) algorithm was proposed for nonparametric maximum likelihood estimation (NPMLE) of the Cox model with interval-censored data, that allows a non-parametric event-time distribution and produces consistent, asymptotically normal and asymptotically efficient estimators for regression parameters [133]. The EM-algorithm is valuable in numerical optimisation when there is incomplete or hidden data. In the interval-censored approach, the parameters can be obtained by maximization of the function given in **Equation 4** [134], however direct maximization is challenging due to insufficient information provided by observed data such as lack of adequate coverage of intervals, and the absence of an analytical expression for the hazard rate  $h_k$ .

**Equation 4. Likelihood function to be maximized in interval-censored cox model**

$$\prod_{i=1}^n \exp \left\{ - \sum_{t_k \leq t_{li}} h_k \exp (x_i(t_k) \beta) \right\} \left[ 1 - \exp \left\{ - \sum_{t_{li} < t_k \leq t_{ui}} h_k \exp (x_i(t_k) \beta) \right\} \right]^{I(t_{ui} < \infty)}$$

$x_i(t_k)$  are the covariate values for subject  $i$  at time  $t_k$ , with  $k=1, \dots, m$

$t_{li}$  is the lower time point for subject  $i$ , and  $t_{ui}$  is the upper time point for subject  $i$

$I(t_{ui} < \infty)$  used to denote an indicator function which takes value 1 if the parentheses evaluates to true.

Zeng (2016) et al [135] proposed the EM-algorithm, where they constructed latent Poisson variables that overcame issues of direct maximization of **Equation 4**. This would be equivalent to maximizing (through the EM-algorithm) the below presented likelihood for the observed data when  $\sum_{t_k \leq t_{li}} W_{ik} = 0$  and  $I(t_{ui} < \infty) \sum_{t_{li} < t_k \leq t_{ui}} W_{ik} > 0$  treating  $W_{ik}$  as missing data (**Equation 5**) [134], however this interval-censored algorithm is computationally intensive to run.

**Equation 5. Interval-censored likelihood for observed data with independent latent Poisson random variables for the EM-algorithm**

$$\prod_{i=1}^n \prod_{t_k \leq t_{li}} P(W_{ik} = 0) \left\{ 1 - P \left( \sum_{t_{li} < t_k \leq t_{ui}} W_{ik} = 0 \right) \right\}^{I(t_{ui} < \infty)}$$

$W_{ik}$ , with  $i=1, \dots, n$ ;  $k=1, \dots, m$  are independent latent Poisson random variables with means  $h_k \exp (x_i(t_k) \beta)$ . It is assumed that  $\sum_{t_k \leq t_{li}} W_{ik} = 0$  and  $I(t_{ui} < \infty) \sum_{t_{li} < t_k \leq t_{ui}} W_{ik} > 0$ .

More recently, a paper by Steyerberg outlined 7 steps to consider in development and validation studies [136]. **Appendix 3** shows how **P7** has addressed each step.

## Results

The development cohort included a total of 40,334 participants. Multivariable hazard ratios of the derived models are shown in **Table 11**, including the baseline survival values at 3-years, needed for specification of the risk model equations. An interaction between age and duration of diabetes was found to improve model discriminatory ability, and therefore included in all derived models.



**Table 11. Hazard ratios for risk models predicting three-year risk of STDR**

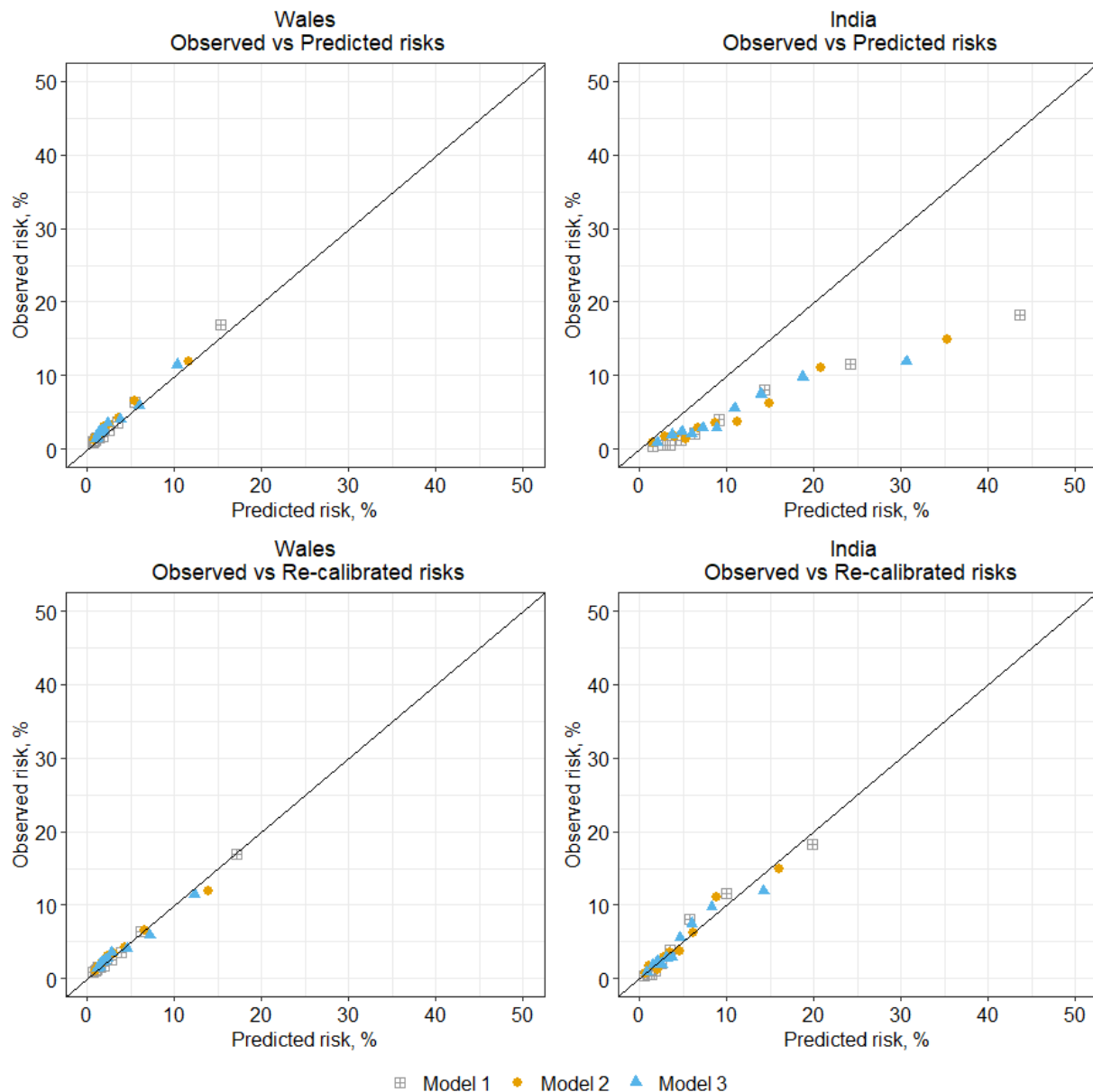
Characteristic	Model 1 (N=40,334)	Model 2 (N=40,334)	Model 3 (N=40,334)
	Hazard Ratio (95% CI)	Hazard Ratio (95% CI)	Hazard Ratio (95% CI)
<b>Age</b>			
<45	1.00	1.00	1.00
45-54	1.17(0.91-1.50)	1.21(0.94-1.55)	1.16 (0.91-1.49)
55-64	1.16(0.89-1.51)	1.25(0.97-1.63)	1.11 (0.86-1.44)
65-74	1.65(1.26-2.15)	1.70(1.31-2.22)	1.44 (1.11-1.88)
75+	1.80(1.27-2.53)	1.93(1.38-2.70)	1.56 (1.12-2.19)
<b>Duration of Type 2 Diabetes (Years)<sup>a</sup></b>	1.09(1.06-1.11)	1.12(1.09-1.15)	1.12(1.10-1.15)
<b>Age by duration interaction<sup>a b</sup></b>			
<45	1.00	1.00	1.00
45-54	0.99(0.96-1.01)	0.98(0.95-1.00)	0.98(0.95-1.01)
55-64	0.97(0.95-1.00)	0.96(0.93-0.98)	0.96(0.93-0.99)
65-74	0.96(0.93-0.98)	0.94(0.91-0.96)	0.94(0.91-0.96)
75+	0.95(0.92-0.97)	0.93(0.90-0.95)	0.93(0.90-0.95)
<b>Gender</b>			
Male	1.00	1.00	1.00
Female	0.89(0.80-0.99)	0.84(0.76-0.94)	0.83(0.75-0.92)
<b>Antidiabetic History</b>			
Diet control	1.00	1.00	1.00
One drug	1.35(1.05-1.73)	1.37(1.07-1.75)	1.49(1.16-1.90)
Two drugs	2.42(1.91-3.07)	2.74(2.16-3.47)	3.55(2.81-4.48)
Insulin	3.43(2.66-4.42)	4.45(3.46-5.72)	6.75(5.29-8.62)
<b>Hba1c (mmol/mol)</b>			
<50	1.00	1.00	
50-59	1.19(0.98-1.44)	1.23 (1.02-1.49)	
60-69	1.69(1.39-2.05)	1.80 (1.49-2.19)	
70-79	1.82(1.47-2.25)	2.03 (1.64-2.50)	
80 and over	2.88(2.39-3.46)	3.28(2.73-3.93)	
<b>History of Background (mild or moderate) DR</b>			
No	1.00		
Yes	3.71(3.30-4.16)		

Abbreviations: DR-diabetic Retinopathy; STDR-Sight threatening diabetic retinopathy; CI-confidence interval. <sup>a</sup> Continuous duration of type 2 diabetes was modelled. <sup>b</sup> Age by duration interaction effect. Shrunk (Heuristic) baseline survival at 3-years in model development dataset is 0.9947 for model 1, 0.9933 for model 2 and 0.9903 for model 3. Extracted from P7[62].

Internal validation yielded c-statistics of 0.832 for model 1 consisting of age, gender, duration of T2DM, age-duration interaction, antidiabetic drugs, HbA1c and background DR, 0.795 when excluding presence of background DR, and 0.778 excluding both background DR and HbA1c.

External validation, performed on 102,672 participants from Wales (SAIL) and 17,509 participants from Chennai, India (MDRF), yielded c-statistics ranging 0.685-0.823 and calibration slopes closer 1 following model re-calibration (**Figure 8**). Risk charts for the Model 3 provide estimates for 3-year risk of STDR, used to aid community workers to prioritise patients for retinal screening, in resource restricted settings (**Figure 9**). Sensitivity analysis confirmed that there was negligible difference in the coefficients from Cox regression and the interval-censored cox models, and showed close correspondence between STDR incidence rates using the interval-censored Cox model and Kaplan-Meier rates in the development cohort.

**Figure 8. Calibration plots for model's 1, 2 and 3 showing observed vs predicted 3-year risk of STDR in validation cohorts**



*Abbreviations: STDR-Sight threatening diabetic retinopathy. Model 3 is the non-invasive model. Risks were categorized into deciles of predicted risk in two external validation datasets (UK-SAIL and India-MDRF). Shrunk (Heuristic) baseline survival at 3-years was used to generate predicted risks in the external validation cohorts. Extracted from P7 [62].*

**Figure 9. Risk-chart showing 3-year % risk of STDR using the non-invasive model (Model 3) in model development cohort**

T2DM Duration *	Age	Male				Female			
		No Antidiabetes	One Drug	Two Drug	Insulin	No Antidiabetes	One Drug	Two Drug	Insulin
0 to <5 Years	<45	1%	2%	5%	8%	1%	2%	4%	7%
	45-54	1%	2%	5%	9%	1%	2%	4%	8%
	55-64	1%	2%	5%	8%	1%	2%	4%	7%
	65-74	2%	2%	6%	10%	1%	2%	5%	9%
	75+	2%	2%	6%	11%	1%	2%	5%	9%
5 to <10 Years	<45	2%	3%	8%	15%	2%	3%	7%	12%
	45-54	2%	3%	8%	14%	2%	3%	7%	12%
	55-64	2%	3%	6%	12%	2%	2%	5%	10%
	65-74	2%	3%	7%	13%	2%	3%	6%	11%
	75+	2%	3%	7%	13%	2%	2%	6%	11%
10 to <15 Years	<45	4%	6%	14%	24%	3%	5%	11%	21%
	45-54	4%	5%	12%	22%	3%	4%	10%	18%
	55-64	3%	4%	9%	17%	2%	3%	8%	14%
	65-74	3%	4%	9%	17%	2%	3%	8%	14%
	75+	2%	4%	8%	15%	2%	3%	7%	13%
15 to <20 Years	<45	7%	10%	23%	39%	6%	9%	20%	34%
	45-54	6%	8%	19%	32%	5%	7%	16%	28%
	55-64	4%	6%	13%	24%	3%	5%	11%	20%
	65-74	3%	5%	12%	21%	3%	4%	10%	18%
	75+	3%	4%	10%	18%	2%	4%	8%	15%
20 + Years	<45	12%	18%	37%	59%	10%	15%	32%	52%
	45-54	9%	13%	28%	46%	7%	11%	24%	40%
	55-64	6%	8%	19%	32%	5%	7%	16%	28%
	65-74	4%	7%	15%	26%	4%	5%	13%	22%
	75+	4%	5%	12%	22%	3%	4%	10%	18%

Abbreviations: T2DM- Type 2 Diabetes Mellitus. Extracted from P7 [62]. Colour schema: Green-Low risk, Orange-Medium risk, Red-High risk.

## Originality and Contribution to the subject

There is a global need to shift towards personalised approaches to predict STDR due to the rising prevalence of diabetes. This publication presents the first non-invasive predictive model for STDR in people with T2DM. Most existing models suffer from bias, due to small sample sizes, missing data, or lack of external validation. A user-friendly risk chart was created with colour-coded risk levels for improved risk interpretations. Similar models like the QRISK tool for heart disease are already used by GPs [137]. For STDR, an individualized risk assessment tool has been developed by the Icelandic group [54] with low risk of bias (minimal missing data, multiple external validations and adequate sample size) and incorporates few laboratory tests or clinical examination parameters. However, these models require retinal images and HbA1c data. **P7**'s key contribution is predicting disease progression without these requirements. The model's achieved good discrimination and calibration of risks even without HbA1c and retinopathy data. It also presents resource-driven models adaptable to different settings, addressing acute global shortages in eye care personnel and medical infrastructure facilitating a transition from opportunistic to need-based DR screening.

## Critical reflection

The availability of the interval-censored cox model in Stata 17 [109], as well as feedback during peer review enabled us to incorporate interval-censored methods into our study. This study's small proportion of interval-censored outcomes (i.e., those who had experienced an event within the duration of the study) had minimal impact on model coefficients compared to Cox regression. I proposed a method to reduce biases when defining time to STDR; both in the definition of censoring and in the modelling process. EHR data, being rich with structure means that it requires detailed statistical considerations. This required some careful planning and preparation of the data to maximise the use of the available data to ensure incidence rates in the population were accurately quantified. The study represented an important step to wider access, however further testing in local datasets will be required for local calibration, to address the variation in prevalence of STDR and ethnicity.

In summary, **P7** presents three resource-driven STDR prediction models; of which the least resource exhaustive model can predict progression to STDR without retinal images or laboratory parameters, making it an ideal choice for use in poorly resourced settings.

## Chapter 6. Prognostic modelling in CKD using routine EHR data

### Context and Objective

Similar to prognostic modelling research in DR, CKD prediction models for population-based screening are not currently useable in resource restricted settings. Despite the ubiquity of CKD prediction models [138], CKD prediction is a developing field and in need of more investigations to improve the adoption of risk models for decision making in CKD prevention strategies. Further investigations are needed to identify individuals at risk of CKD so that clinicians can implement prevention strategies to reduce the likelihood of comorbidities of T2DM. The findings from **chapter 2** consolidate existing links between DR and DKD. DR is a known comorbidity in people with CKD, and therefore also found simultaneously in individuals experiencing comorbidities of T2DM. In our studies, impaired renal function was a predictor for the development of DR and STDR (**P4, P5 and P6**).

This article considers the utility of incorporating routine clinical parameters used in our DR prognostic model (**P7**), to derive clinically usable risk models for CKD (**Table 12**). It also examines risk factors for CKD.

**Table 12. Chapter 6 Publication 8, citations and mentions, updated 16/03/2024.**

P8: CKD risk prediction modelling study

Development and validation of resource-driven risk prediction models for incident chronic kidney disease in type 2 diabetes

**Gurudas S\*, Nugawela M\***, Prevost AT, Sathish T, Mathur R, Rafferty JM, Blighe K, Rajalakshmi R, Mohan AR, Saravanan J, Majeed A, Mohan V, Owens DR, Robson J, Sivaprasad S

2021, Scientific reports, Impact factor: 4.011



### Methodological commentary and Critical powers

The development and validation process in this study mirrored that of **P7**; closely following the TRIPOD guidelines and recommendations by the PROGRESS framework [131, 132].

The outcome definition relied on both laboratory tests for eGFR and clinical codes for diagnosing stage 3+ CKD, ensuring robust case identification and reducing the risk of underestimation of

incidence. Additionally, stage 3+ CKD diagnosis required two eGFR measurements below 60 ml/min/1.73m<sup>2</sup>, at least 3 months apart with the earlier of the two measurements being considered as onset of disease.

Four measures were proposed to evaluate model performance in Steyerberg's paper [136]; (A) *calibration in the large or the model intercept*; (B) *calibration slope*; (C) *discrimination*; and (D) *clinical usefulness with decision-curve analysis*, all of which were considered in our study, graphically and/or quantitatively. Discrimination was assessed using Harell's c-statistic. Calibration was assessed using the beta coefficient of linear predictor (LP) and observed to expected ratio (O/E). Calibration slopes were also visually examined in several clinically relevant subgroups. **Appendix 3** shows how **P8** has addressed each of the 7 steps for risk prediction development and validation.

Continuous variables that exhibited non-linear risk relationships were modelled as continuous variables using fractional polynomials [139-141] and interactions between variables were assessed based on clinical literature and apriori knowledge. eGFR and ACR were included in the final models despite being invasive tests due to their importance in CKD modelling research. The full model consisted of all variables that were identified from backward elimination, including demographic, laboratory, medication history, cardiovascular disease history (CVD) and STDR. The second model was developed by excluding CVD and STDR (reduced model) and the third model developed by excluding HbA1c and High-Density Lipoprotein (HDL) (minimal resources model). The least resource intense model was presented as a risk score for ease of use in resource restricted settings.

Cox proportional hazards models were used for statistical modelling (**Equation 3**). Furthermore, missing at random (MAR) assumptions were assessed, as Cox regression, fit via maximum likelihood on complete cases can introduce bias when the data are not MAR. Missing data for microalbuminuria affected about 50% of the UK cohorts, making it a significant challenge.

Novel decision curve analysis (DCA) [142, 143], a method for evaluating and comparing prediction models incorporating clinical consequences was used to assess the utility of the derived risk models (**Figure 10**). DCA calculates clinical "net benefit" (**Equation 6**) of a clinical model across a range of thresholds and compares it to the decision of treating all or no patients. It's defined as the minimum probability of disease at which the risk model or test should be applied, as it incorporates the harms (unnecessary testing) and benefits (benefit of early detection) of the clinical decision unlike performance metrics that evaluate discrimination and calibration. Risk

models were presented using a points score system to aid implementation in clinical settings [144].

**Equation 6. Formula for the net-benefit of a risk model at a given threshold.**

$$Net\ benefit = \left( \frac{True\ positive\ count}{N} \right) - \left( \frac{False\ positive\ count}{N} \right) * (p_t / (1 - p_t))$$

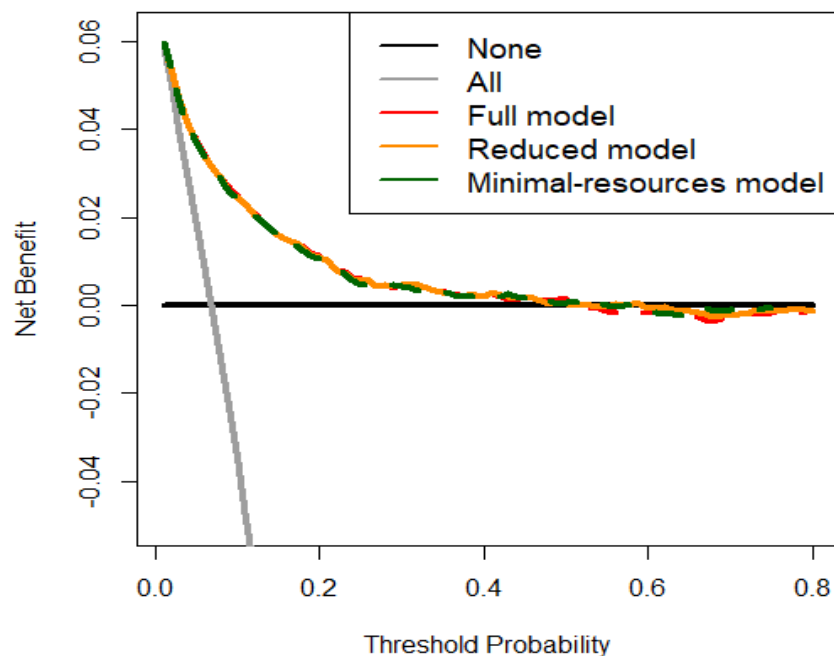
where  $p_t$  is the threshold probability that a patient is positive based on the risk model

## Results

The models were developed on a large ethnically diverse cohort of primary care registered individuals from inner London (used in **P4** and **P7**). In total, 20,510 (East London dataset) were used for development and 13,346 for validation (SAIL). Baseline characteristics of participants with missing data were not significantly different from complete data. Age was found to be best modelled using fractional polynomials and with an interaction effect with use of insulin ( $p < 0.001$ ) in all three models. STDR was identified as a statistically significant predictor for incident stage 3+ CKD. All models achieved good accuracy, with Harell's c-statistics ranging 0.852-0.853 in internal validation and 0.823-0.827 in external validation. The beta-coefficient of the LP was near to 1 in internal validation and ranged 1.02-1.03 across three models in external validation. Predicted risks were better aligned with observed risks following re-calibration of the baseline survival estimate at 5-years in the Wales cohort. Model 3 can stratify risk of patients with normal eGFR ( $eGFR \geq 90\text{ ml/min/1.73m}^2$ ) to a range of thresholds (0%-46%) at 5 years. All three models had similar net benefit (**Figure 10**), with the minimal resources model (model 3) identifying 6 more cases of CKD per 100 screened without increasing the number treated unnecessarily at a threshold of 10%. **Figure 11** presents the mapping of points to predicted probabilities for the minimal resources model, for both UK cohorts.



**Figure 10. Decision curves comparing CKD models in the external validation cohort**



*This graph shows the expected net benefit for each threshold probability for the 5-year risk of incident CKD evaluated from 0 to 80% relative to screening no one in the population. “None” or screening no one in the population (Black line). “All” or screening everyone in the population (Grey line).*

*“Full model” (Red line), “Reduced model” (Orange line), “Minimal-resources model” (Green line).*

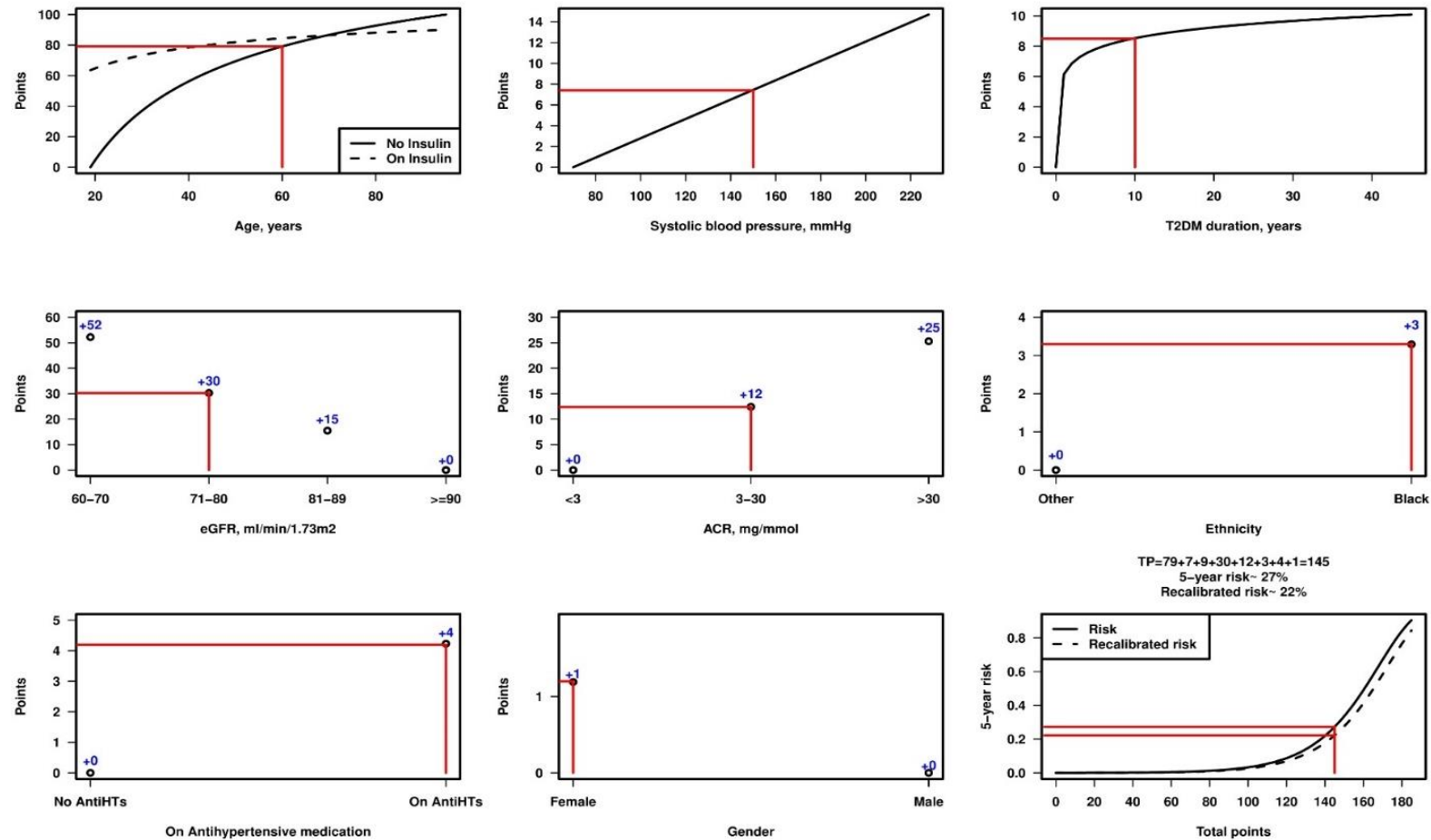
*Net benefit is defined by  $net\ benefit = \left( \frac{True\ positive\ count}{N} \right) - \left( \frac{False\ positive\ count}{N} \right) * (pt / (1 - pt))$ ;*

*where  $p_t$  is the probability threshold. True positive count and false positive count defined by*

*# True positives =  $[1 - s(t)|z = 1] * P(z = 1) * n$  and # False positives =  $(s(t)|z = 1) * P(z = 1) * n$ ,*

*where  $s(t)$  is the survival probability at time  $t$ ,  $z$  is an indicator variable taking value 1 if the predicted probability for the patient  $\geq p_t$ . Extracted from P8 [63].*

**Figure 11. Risk score interpretation for the prediction of 5-year risk of stage 3 + CKD (minimal resources model)**



Abbreviations: T2DM- Type 2 Diabetes Mellitus, eGFR- estimated Glomerular Filtration Fate, AntiHT-antihypertensive, TP-Total Points.  
 Example patient: Age 60 years, SBP 150 mmHg, duration 10 years, eGFR 75 ml/min/1.73m<sup>2</sup>, ACR 15 mg/mmol, Black ethnicity, On AntiHTs and Female gives total points of 79+7+9+30+12+3+4+1=145 points equivalent to 5-year risk of 27% and recalibrated 22%.  
 Extracted from P8 [63].

## Originality and Contribution to the subject

This is the first publication to derive resource-driven risk models for CKD in T2DM to aid decision making in CKD prevention, emphasising fewer laboratory parameters, addressing resource limitations in LMICs.

**P8** contributes clinically usable CKD risk models, incorporating routine clinical parameters collected in the diabetes clinic, employing a range of sound statistical methods, and following most up to standard guidelines on risk prediction modelling. Moreover, the successful external validation and calibration of the models to a second UK cohort (Wales) provides further evidence that the models are stratifying individuals by their level of risk with good accuracy and providing accurate estimates of risk in the population. This study also laid the groundwork for future external validation studies to test model transportability to new populations.

## Critical reflection

The results presented in this thesis reinforce the importance of using routine data to enhance the detection of diabetes complications such as DR and DKD. This study explored methodological issues not explored in current CKD prediction modelling studies, including non-linearity of covariates and decision curve analysis. Since its publication in 2021, **P8** has been cited 4 times. Recent studies stress the need for cost-effective CKD risk prediction models in LMICs, emphasizing the practicality of using markers available in LMIC diabetes clinics [138]. We excluded HbA1c, HDL, CVD and STDR from the full models for two reasons; their limited contribution to the model's c-statistic and the ease of administering these tests in clinical settings. eGFR and urine albumin were retained in models due to their importance in CKD modelling, excluding them could mis-calibrate risks and substantially reduce model discrimination. However, policymakers must be persuaded of the viability of including laboratory markers not routinely collected in LMICs, particularly urine tests with high missing data. Moreover, laboratory testing may incur additional costs, such as handling, processing and analysing blood samples, and may not be easily administered in community screening in LMICs. Alternative, non-laboratory features may need to be modelled, to improve the adoption of risk models into clinical practice in resource-poor settings. Future studies should aim to develop an office-based CKD risk score that balances accuracy and utility, so that these models can be applicable to populations living in different environmental conditions.

In summary, this study presents three economically viable risk models for risk stratification of future stage 3+ CKD, even amongst individuals with normal kidney function. **P7** and **P8** highlight the predictive potential of routine data from diabetes clinics for both STDR and CKD.

## Chapter 7. Synthesis

### Contextualisation of research, Impact, and Study strengths

This section aims to provide a contextualisation of the research within the portfolio and the broader scientific landscape. It explores the potential impact of our findings on the field, the links between the studies included in the portfolio (**Appendix 1**), as well as highlight the notable strengths and contributions of our study.

The 8 publications within the portfolio together support the development of low-cost DR detection tools to overcome the implementation challenges in LMICs in response to the diabetes epidemic. The aim of this portfolio was to illuminate the global challenges of DR detection, identify its key risk factors, the limitations of existing strategies and identify resource-driven screening solutions to reduce the global burden of DR. The research contributed to a major policy change in the Indian state of Kerala [86], where DR screening is now recommended in the guidelines for PwD, to be implemented in all family health centers across the state. This major achievement can be attributed to several reasons, some of which relate to the work contained in the portfolio:

- i) A more robust understanding of the global literature on DR prevalence, current practices, and gaps in DR research (**chapter 1**)

**P1** motivates the need for this research as it highlights the burden of disease globally. It introduces the grades of DR in varying severity, provides a general overview of the current literature on DR prevalence, and attempts to synthesise the evidence by various epidemiologically relevant subgroups, using robust inclusion criteria. **P1** concludes by highlighting the challenges faced when synthesising evidence on the prevalence of DR due to heterogeneity in the included studies. These features of study design were carefully considered across all studies within this portfolio.

Moreover, the lack of studies in non-European countries and resource-poor settings highlight the need for more global coverage in DR screening, especially in ethnic minority groups, including Indians and this was the focus in the remaining chapters of this thesis.

A secondary conclusion was the need for alternative pathways for screening. The rising prevalence of STDR, yet its relatively small proportion (~0.6% PDR and ~1.3% DMO) in T2DM, makes screening all people with T2DM inefficient. While LMICs are harder hit by the consequences of diabetes such as blindness, due to population ageing, VI and blindness are an

increasing economic burden in high-income countries (HICs). The publications in this thesis aim to cater to all resource settings.

- ii) Robust assessment of the prevalence and incidence of VI and blindness in India due to DR (**chapter 2**)

**P1** assessed the global screening burden and highlighted the lack of studies conducted in LMICs; **P2** and **P3** attempted to fill this gap by providing quantitative estimates of the burden of VI in DR. **P2** also provided estimates of VI and blindness with national coverage in PwD. Complex survey sampling was used in **P2** to recruit an adequate sample size, and in the analysis, newly derived weights were used to ensure national representativeness of our findings. Primary data collection by the SMART-India study collaborators ensured efficient data collection, as complex survey sampling can reduce costs compared to a complete enumeration of the population. These methods also allow researchers to make inferences that account for the surveys design features.

- iii) Comprehensive understanding of the risk factors for incidence of DR and the role of ethnicity with an interest in UK Indians (**chapter 3**)

A key finding from **P1** was the global, regional, and ethnic variations in DR rates. **P4** systematically investigated this hypothesis within an ethnically diverse UK cohort and identified the key risk factors associated with incident DR. The identification of these risk factors underscores the importance of implementing stratified or risk-based screening approaches in diabetes management programs. It concludes by identifying that ethnic minorities are at increased risk of DR and STDR compared to their white counterparts, independent of lipid profile, blood glucose control, duration of diabetes, kidney function and diabetes medication-use. These results highlight that ethnic minorities in a UK sample, who are invited for systematic screening are still at increased risk for incident DR and STDR and should be targeted for care. Both parametric and semi-parametric cox models were considered for statistical modelling. However, due to the flexibility, robustness and interpretability offered by cox models, was selected as a reasonable approximation for modelling the time to event outcome.

- iv) Identification of alternative pathways to DR screening using biomarkers, that can aid and reduce backlog in systematic screening of DR in India and UK (**chapter 4**)

**P5** corroborates the evidence that blood glucose control should be prioritised in treatment programs, and most importantly ranks HbA1c as the strongest ranking marker among over 400 markers assessed. NHANES has several design features including probability sampling,

stratification, cluster sampling within each stratum and oversampling of ethnic minorities and special populations, which help to capture the diverse U.S. population. Analytical considerations in this study align with those used in **P2**. A key limitation for **P5** was the lack of assessment of non-routine laboratory markers for STDR, known in the literature to be distinguished in people with STDR. This motivated the need for a theory-driven data analysis, where various mechanisms of action were hypothesised, prompting the collection of original data in **P6** to assess these mechanisms.

**P6** uses a confirmatory approach, with scientific theory guiding the selection of variables. The research design differs from that of **P5**, as it involves primary data collection to create the analytic dataset. It also provides evidence for biomarkers identified in prior research, not gathered in **P5**. Again, weighting was employed, with intentional oversampling of outcome groups, resembling a case-control design with three outcome groups. This efficient approach helped reduce data collection costs and facilitated the collection of data on 13 non-routine biomarkers. Furthermore, it spared the need to recruit a large number of observations to achieve the desired number of events for modelling. Serum creatinine and Cystatin-C, markers of kidney impairment, could stratify STDR risk in both publications, further supporting causal links established in **P4's** longitudinal analysis. Additionally, replicating the study in two independent cohorts, using consistent standards, enhanced the robustness of our findings. The findings of **P6** were used to support the development of a biosensor incorporating Cystatin-C, a project currently underway in a hospital in south India.

- v) Non-invasive and resource-driven screening solutions with good accuracy to aid the prognostication in resource-poor settings such as community screening in India (**chapter 5**)

**P7** is central to the portfolio and a culmination of three years of work contributing to WP5 of the Ornate India Project. WP5 aimed to use big data and population level databases to develop and validate risk models for diabetes complications, particularly DR and STDR, for use in LMICs. A non-invasive, cost-effective risk tool can facilitate patient prioritisation and population-level risk stratification, addressing the backlog in DR screening.

**P1** quantified the global DR burden and emphasised the need for LMICs to implement screening programs, motivating **P7**. **P3** highlighted the extent of VI and blindness in PDR patients despite treatment, again emphasizing the importance of early detection and timely treatment. The variables identified in **P4** were evaluated for their practicality, cost, and ease of administration,

then translated into risk equations supplemented through a risk chart in **P7**. Age, gender, diabetes duration and anti-diabetic medication-use, assessed in **P4** emerged as robust risk factors for incident DR and STDR. HbA1c and DR status, deemed less cost-effective, were excluded to reduce the burden of screening, findings echoed in **P1**. **P5** and **P6** contributed to the diagnostic criteria for DR and/or STDR, while **P7** focused on STDR prediction in those without it. Methods used in **P7** varied from the other studies as the primary aim was for developing risk prediction models, this included the need to validate derived risk equations in multiple cohorts. Unlike the other studies, it did not focus on explanation or addressing confounding but prioritized the utility of modeling with fewer variables, assessing the discrimination and calibration it achieved.

- vi) Continued applications of resource-driven risk modelling solutions using routinely collected laboratory variables for the prediction of incident CKD in T2DM (**chapter 6**)

**P8** utilised routinely collected EHR data to predict CKD in individuals with T2DM and developed risk models using routinely collected laboratory markers. Laboratory variables such as eGFR and ACR, although not non-invasive, can be used in guiding prediction of CKD due to its relatively cheap cost and predictive accuracy for incident CKD. While not all individuals with low eGFR and high ACR develop CKD, these tools model the complex relationship between eGFR, ACR, age, gender, and duration of diabetes to be used in clinical practice, alongside evidence of additional risk factors. Methods used in this publication adhered to TRIPOD guidelines and PROGRESS framework, including the novel DCA approach, as well as use of fractional polynomials to model non-linear risk relationships, contributing to the study's robustness and exemplifying good analytical practices. **P5**, **P6** and **P8** collectively highlight the correlation and bi-directional relationship between CKD and DR in people with T2DM, underscoring the importance of simultaneous screening for both conditions.

## Limitations

Details on limitations of each study were given within each respective chapter. A key limitation in this thesis is a lack of investigations into genetic determinants of DR. As the UK health systems move towards digital health, personalization and real-time monitoring of risk factors can alter the usability of risk models in the NHS. The role of genetics in DR can contribute to achieving the 4 P's as set out by NHS England (**P**rediction and Prevention, More **P**recise diagnoses, **P**ersonalised and targeted interventions, and a more **P**articipatory role for patients). The genetic variants in DR are yet to be elucidated and replicated in confirmatory analysis which would help clinicians tailor



treatments to patients with DR [145]. However, this should be supplemented with lifestyle modifications, informed by the research undertaken in this thesis.

## **Contribution to the field, Implementation challenges and Future work**

Key implementation challenges of risk models into clinical practice include; i) the resistance from staff and patients if risk models were to replace the current standard particularly in LMICs where risk tools are nascent, ii) the potential mistranslation of risk tools used in practice, as methods to rule out patients from DR screening, and its hindrance on model uptake and iii) the lack of evaluation of cost-effectiveness of clinical prioritisation methods and the measurable reduction in screening backlog.

The current pathway for the DESP in the UK is 1) screening asymptomatic individuals with diabetes to identify those at risk of DR, 2) diagnosis of the disease and 3) treatment if required. The national unit average for Diabetic Eye Disease screening in the UK is £29 per person according to 2014-15 estimates [146]. While annual screening may not be feasible when the screening burden is 5 million PwD in the UK, it is widely agreed that targeted screening can improve service organisation and help tackle the elective backlog. For instance, it has been proposed that the NHS should continue to focus on clinical need as they tackle the backlog from COVID-19 [147]. Such methods are already in use for cancer referrals in the UK. In this new model of care, waiting lists could be managed in order of priority based on risk factors including age, gender, diabetes duration, ethnicity, or laboratory tests, depending on the setting. Detailed action plans would need to be developed following a comparative cost-effectiveness analysis in various settings. Precision medicine advocates for a more individualized strategy in addressing care. The NICE guidelines under *1.1.1. individualised care* states individualised care that is “tailored to the needs and circumstances of adults with type 2 diabetes, taking into account their personal preferences, comorbidities and risks from polypharmacy, and their likelihood of benefiting from long-term interventions” [148] is needed.

In the UK, ongoing efforts to enhance cost-effectiveness of the diabetic eye screening program include considering biennial or variable screening intervals [149] [150] [7] [54] [151], with proposals for extending intervals to 1 to 2 years for people without retinopathy on consecutive screens. Despite these proposals, there is a consensus that systematic screening is cost-effective compared with opportunistic screening [152]. In England, the screening programme has relegated DR from being the leading cause of certifiable blindness in working-age individuals [153]. These programs aim to detect DR before less treatable microvascular changes occur. Therefore,

incorporating an additional step prior to retinal photography to identify individuals at high-risk that need to be prioritised in the sequence of screening, diagnosis, and treatment will alleviate the screening burden and facilitate early diagnosis.

Efficient pre-screening tools should reduce over-screening costs without forgoing sensitivity. In **P5**, diagnostic models with Cystatin-C demonstrated good sensitivity and specificity, inspiring an ongoing study in India to develop a Cystatin-C biosensor, enabling a transition from traditional analytical off-site laboratories, offering substantial cost savings for governments and easing the burden on healthcare systems [154]. Rigorous and extensive replications of this study are needed to assess thresholds of Cystatin-C for detecting STDR [155], such as a recent application of Cystatin-C in Asian Indians which found significantly higher Cystatin-C levels in STDR compared to no STDR and no DR [156]. Biosensors are increasingly used for screening infectious diseases, early detection and managing well-being, particularly wearable biosensors [157, 158]. Advances in nanotechnologies have driven these innovations, although rigorous testing across different settings remains crucial.

**P6's** prognostic models, used in people with no DR, can be applied to the least resource intense settings and require little to no startup costs, as the minimal resource model is fully non-invasive. The risk charts are compact and visually attractive, ideal in community screening settings in LMICs. While I conducted 2 external validations, further validations are still needed before clinical use [155]. Clinical prediction models are largely underutilized in health care practices world-wide including the UK, though the usability of any model is greatly reduced if there is a lack of integration with IT [159, 160]. QRISK is a tool that has overcome these challenges, which has been embedded within several primary care management systems in the UK [161]. However, risk charts would be a cost-effective alternative during a transition period for health systems that are facing huge backlog and don't have the resources to support digital healthcare. Models in **P7** and **P8's** models would work well if embedded in EHRs in primary and secondary care settings. However, non-standardised coding practices across EHR providers may be a hindrance to unifying practices across settings.

Thirdly, rigorous assessments of the cost-effectiveness of this alternative screening pathway will need to be modelled, to elucidate both the benefits in terms of cost-savings and harms in terms of cases that are subject to late diagnoses.

The focus on this thesis has been on improving health outcomes for LMICs, building research capacity, and identifying solutions for reducing the global screening burden of DR. While the

prevalence and burden of DR vary in the UK and India, similar conclusions on viability of our proposed strategies were drawn in datasets from both countries.

## **Contributor statements**

Full author contribution statements can be found within each publication and a summary of my contributions to each study activity is provided in **Figure 12**. In **chapter 1**, I reviewed the included studies, collated the data from each study, prepared the tables, acted as an arbiter for disagreements and critiqued the manuscript. I provided support to the conceptualisation, selection of studies, introduction, methodology, results, and discussion sections. In **chapter 2** publications, I was responsible for all aspects of **P2** from conceptualization to analysis, manuscript preparation and submission (currently under review). For **P3**, I produced the revised statistical analysis plan, conducted the analysis, critically reviewed the manuscript, and responded to the statistical queries during peer review. In **chapter 3**, I created the statistical analysis plan, acquired, and preprocessed the primary care data, selected the statistical model, interpreted univariate and multivariable analyses, provided extensive critique to methodology and results and reviewed the manuscript. In **chapter 4** publications, for **P5**, I contributed to the conceptualization, literature review, statistical analysis plan, selection and application of survey weights to the logistic regression model, results interpretation and critical review. For **P6**, I contributed to all sections except for the data collection and study conceptualisation. In **chapter 5 & 6**, I collaborated with my joint co-author on the literature review, protocol development, statistical analysis plan, study design, sample size calculations, dataset curation, candidate variable selection, risk model development, external validation, critical review, submission and co-leading the peer review process.

**Figure 12. Contributorship matrix showing my contributions for P1-P8**

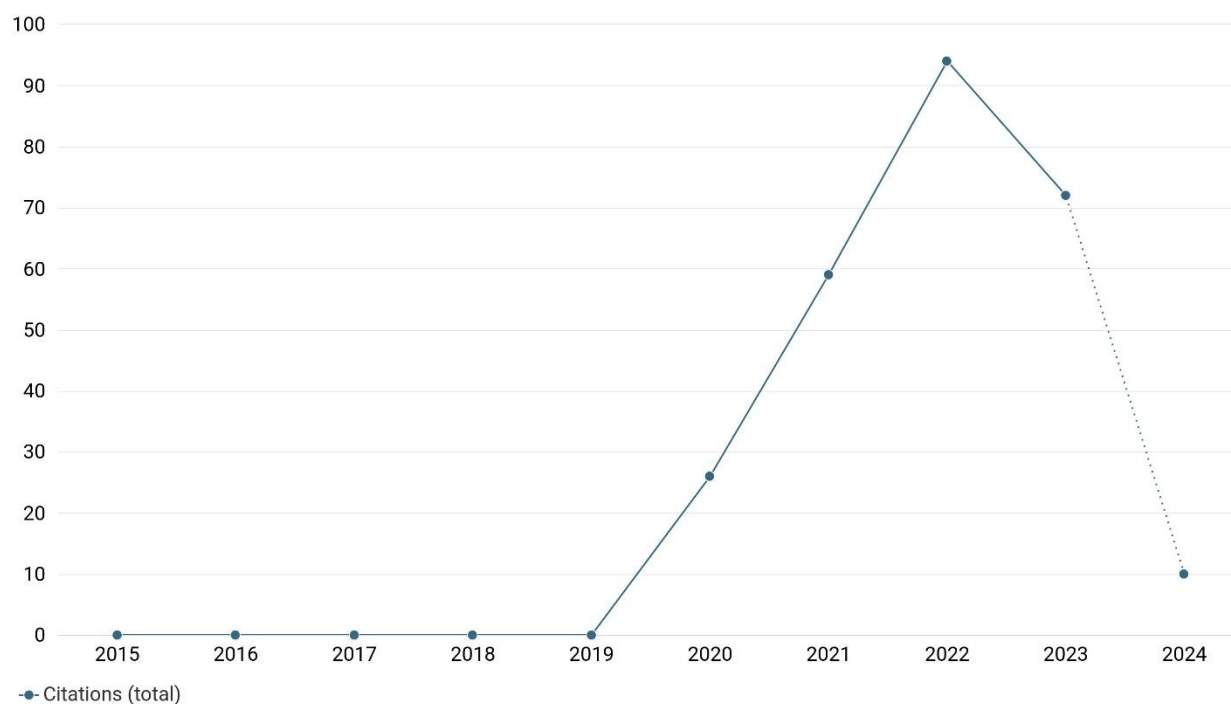
	P1	P2	P3	P4	P5	P6	P7	P8
Conceptualisation	Light Blue	Dark Blue	White	White	Dark Blue	White	White	Dark Blue
Selected studies for literature review	Light Blue	Dark Blue	White	Light Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Statistical analysis plan	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Data preprocessing	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Light Blue	Dark Blue	Dark Blue	Dark Blue
Data analysis	Dark Blue	Dark Blue	Dark Blue	Light Blue	Light Blue	Dark Blue	Dark Blue	Dark Blue
Data interpretation	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Visualisation	White	Dark Blue	Grey	Light Blue	Light Blue	Dark Blue	Dark Blue	Dark Blue
Manuscript preparation: introduction	Light Blue	Dark Blue	White	Light Blue	Light Blue	Dark Blue	Dark Blue	Dark Blue
Manuscript preparation: methods	Light Blue	Dark Blue	Light Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Manuscript preparation: results	Light Blue	Dark Blue	Light Blue	Light Blue	Light Blue	Dark Blue	Dark Blue	Dark Blue
Manuscript preparation: discussion	Light Blue	Dark Blue	Light Blue	Light Blue	Light Blue	Dark Blue	Dark Blue	Dark Blue
Manuscript preparation: critical review	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue	Dark Blue
Peer review	Light Blue	Dark Blue	Light Blue	Light Blue	Light Blue	Dark Blue	Dark Blue	Dark Blue

*White-not contributed, light blue-contributed, dark blue-lead/co-lead, grey/shaded-NA*

## Development and growth as a researcher

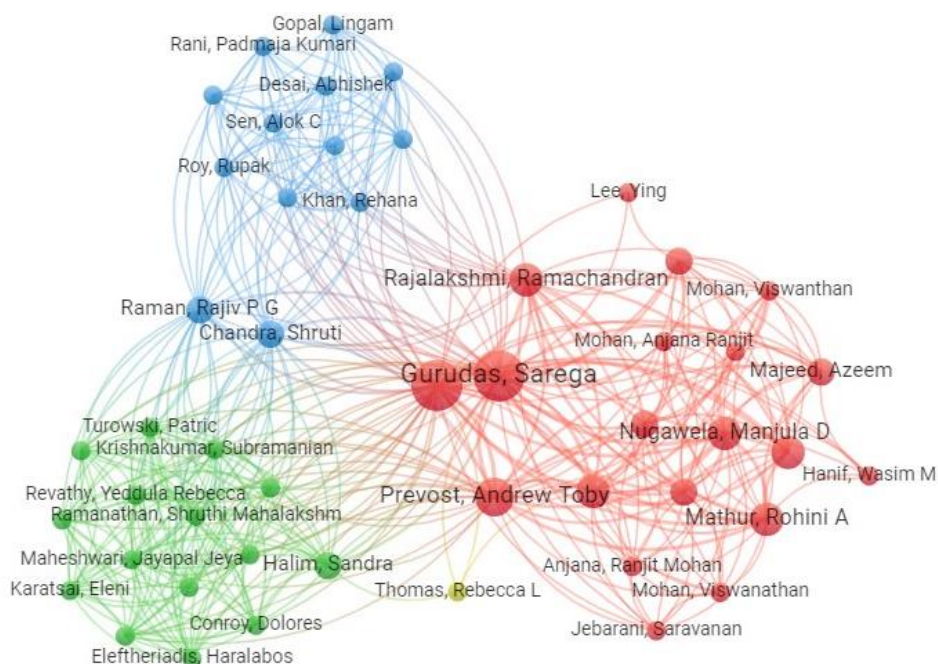
This portfolio signifies my ongoing learning and growth as a researcher throughout the Ornate India project. The contained publications has seen a steady increase in citations each year (**Figure 13**) and **Figure 14** illustrates the expanding network of collaborators behind my publications. Other examples include a poster presentation (of **P2**) at the Coventry University research showcase in April 2023 (winning the poster presentation competition), UCL IoO's annual symposium in June 2023, and an accepted abstract submission and oral presentation at the 2023 IDF virtual congress. These opportunities to network allowed me to receive feedback, learn from other presenters and broaden my professional network.

**Figure 13. Citations received in each year based on P1-P8, updated 16/03/2024.**



*Solid line refers to complete years (2020,2021,2022,2023) and dotted line refers to current year (till March 2024). Extracted from dimensions [162].*

**Figure 14. Network model for research connections formed based on P1-P8**



*Extracted from Dimensions website [162].*



**P1** marked my introduction to the world of publishing. It was also the most cited publication in this thesis.





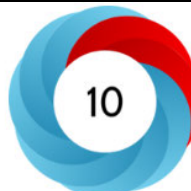



**P2** and **P3** allowed me to understand the challenges of working with data from a different health system, the kind of missing data it presents, differing standards in the criteria used to assess DR as well as the statistical adjustments required in survey research.







**P4**, **P6** and **P7** provided valuable experiences in handling UK EHR data and the challenges of mitigating bias in data not originally collected for research. **P4** and **P7** shared the same dataset but followed different methodologies as **P7** incorporated interval-censored methods in sensitivity analysis due to its availability in Stata 17. **P2**, the final publication, showcased the skills acquired throughout my research journey, from an in-depth literature review in **P1** to leading statistical analysis in **P2**, **P3**, **P6** and **P8** and co-leading in **P4**, **P5** and **P7**, as well as communicating research findings in **P2-P8**.

**P5** was my introduction to machine learning and high-dimensional data analysis with over 400 parameters. The publications were all featured in journals with moderate-high impact factor (2022 IF 3.4-15.1) and continue to attract online attention, with altmetric attention scores ranging from 3-24 as of 16/03/2024 (**Table 13**).

**Table 13. Altmetric and Plumx metrics for P1-P8, updated 16/03/2024.**

Altmetric attention score	Plumx	Publication	Journal and Impact factor
 <div> <p>Picked up by 1 news outlets</p> <p>Referenced in 1 policy sources</p> <p>Posted by 16 X users</p> </div>	 <div> <p><b>Citations</b> Policy Citations: 5 Citation Indexes: 195</p> <p><b>Captures</b> Readers: 383</p> <p><b>Mentions</b> News Mentions: 1</p> </div>	<p>P1 – IDF Diabetes Atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018.</p>	<p><i>Diabetes research and clinical practice</i></p> <p>Impact factor (2022): 5.1</p>

 <p>19</p> <p>Picked up by 1 news outlets Posted by 9 X users</p>	 <p>PLUM</p> <p>Citations Citation Indexes: 1</p>	<p>P2 – National prevalence of vision impairment and blindness and associated risk factors in adults aged 40 years or older with known or undiagnosed diabetes: results from the SMART-India cross-sectional study</p>	<p><i>Lancet Global Health</i></p> <p>Impact factor (2022): 34.3</p>
 <p>4</p> <p>Posted by 6 X users</p>	 <p>PLUM</p> <p>Citations Citation Indexes: 11</p> <p>Captures Readers: 34</p>	<p>P3 – Prevalence and incidence of visual impairment in patients with proliferative diabetic retinopathy in India</p>	<p><i>Scientific reports</i></p> <p>Impact factor (2022): 4.6</p>
 <p>10</p> <p>Picked up by 1 news outlets Posted by 4 X users</p>	 <p>PLUM</p> <p>Citations Citation Indexes: 7</p> <p>Captures Readers: 29</p>	<p>P4 – Ethnic Disparities in the Development of Sight-Threatening Diabetic Retinopathy in a UK Multi-Ethnic Population with Diabetes: An Observational Cohort Study</p>	<p><i>Journal of personalised medicine</i></p> <p>Impact factor (2022): 3.4</p>
 <p>3</p> <p>Posted by 4 X users</p>	 <p>PLUM</p> <p>Citations Citation Indexes: 9</p> <p>Captures Readers: 22</p>	<p>P5 -Diabetic Retinopathy Environment-Wide Association Study (EWAS) in NHANES 2005-2008</p>	<p><i>Journal of clinical medicine</i></p> <p>Impact factor (2022): 3.9</p>

 <p>Picked up by 1 news outlets Posted by 19 X users</p>	 <p><b>Citations</b> Citation Indexes: 9 <b>Captures</b> Readers: 24 <b>Mentions</b> News Mentions: 2</p>	<p>P6 - Multicenter Evaluation of Diagnostic Circulating Biomarkers to Detect Sight-Threatening Diabetic Retinopathy</p>	<p><i>JAMA Ophthalmology</i></p> <p>Impact factor (2022): 8.1</p>
 <p>Posted by 3 X users</p>	 <p><b>Citations</b> Citation Indexes: 7 <b>Captures</b> Readers: 29</p>	<p>P7 - Development and validation of predictive risk models for sight threatening diabetic retinopathy in patients with type 2 diabetes to be applied as triage tools in resource limited settings</p>	<p><i>Lancet e-Clinical medicine</i></p> <p>Impact factor (2022): 15.1</p>
 <p>Posted by 3 X users</p>	 <p><b>Citations</b> Citation Indexes: 5 <b>Captures</b> Readers: 31</p>	<p>P8 - Development and validation of resource-driven risk prediction models for incident chronic kidney disease in type 2 diabetes</p>	<p><i>Scientific reports</i></p> <p>Impact factor (2022): 3.9</p>

Abbreviations: IDF – International Diabetes Federation, NHANES- National Health And Nutrition Examination Survey. Altmetric attention score [163] and Plumx [164] scores extracted from Coventry University pure profile[165].

## Conclusion

This doctoral synthesis encompasses a diverse array of studies from the ORNATE-India project addressing various aspects of DR research. It begins with a global DR literature review, progresses to examining the prevalence and incidence of VI in diabetes, investigates risk factors and potential diagnostic biomarkers in DR. Finally, it presents risk prediction models for STDR and DKD using routine data from both developed and developing countries. The studies within



this portfolio have made several notable contributions to knowledge including overcoming a major cost barrier to make early detection of DR more accessible in LMICs. They have provided insights that can be leveraged in the development of innovative DR screening tools, including laboratory markers that can be utilised in a cost-efficient DR biosensor, the DR-DKD association for holistic diabetes care and a non-invasive tool for the prediction of STDR. The studies also inform aetiological investigations, by providing estimates for the global burden of DR and its impact on the problem of vision impairment in LMICs, providing crucial insights to inform the planning and allocation of resources. Alongside highlighting the collective achievements of the project, this synthesis emphasised my contributions and traces my progression as a statistician from a supportive role, to leading my own studies.

This project also represents a successful interdisciplinary collaboration involving statisticians, biologists, and clinicians and stands as a product of strong methodological grounding, rigorous literature review and interdisciplinary learning. The studies presented here offer strategies to reduce DR rates with worldwide applicability and have, ultimately, established the groundwork for advancing DR research in LMICs.

## References

### Uncategorized References

1. Sun, H., et al., *IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045*. Diabetes Res Clin Pract, 2022. **183**: p. 109119.
2. Stumvoll, M., B.J. Goldstein, and T.W. van Haeften, *Type 2 diabetes: principles of pathogenesis and therapy*. The Lancet, 2005. **365**(9467): p. 1333-1346.
3. Todd, J.A., *Etiology of type 1 diabetes*. J Immunity, 2010. **32**(4): p. 457-467.
4. Anjana, R.M., et al., *Metabolic non-communicable disease health report of India: the ICMR-INDIAB national cross-sectional study (ICMR-INDIAB-17)*. The Lancet Diabetes & Endocrinology, 2023. **11**(7): p. 474-489.
5. *IDF Diabetes Atlas*. 9th ed. 2019, Brussels, Belgium: International Diabetes Federation.
6. UN General Assembly Resolution, *Transforming our world: the 2030 Agenda for Sustainable Development*. UN Doc. A/RES/70/1 (September 25, 2015), 2015.
7. Scanlon, P.H., et al., *Development of a cost-effectiveness model for optimisation of the screening interval in diabetic retinopathy screening*. Health Technol Assess, 2015. **19**(74): p. 1-116.
8. Whicher, C.A., S. O'Neill, and R.I.G. Holt, *Diabetes in the UK: 2019*. Diabet Med, 2020. **37**(2): p. 242-247.
9. Jenkins, A.J., et al., *Biomarkers in Diabetic Retinopathy*. Rev Diabet Stud, 2015. **12**(1-2): p. 159-95.
10. Cheung, N., P. Mitchell, and T.Y. Wong, *Diabetic retinopathy*. Lancet, 2010. **376**(9735): p. 124-36.
11. The Royal College of Ophthalmologists, *Diabetic Retinopathy Guidelines*. 2012, Scientific Department, The Royal College of Ophthalmologists London.
12. *Preliminary report on effects of photocoagulation therapy. The Diabetic Retinopathy Study Research Group*. Am J Ophthalmol, 1976. **81**(4): p. 383-96.
13. Gross, J.G., et al., *Five-year outcomes of panretinal photocoagulation vs intravitreal ranibizumab for proliferative diabetic retinopathy: a randomized clinical trial*. JAMA ophthalmology, 2018. **136**(10): p. 1138-1148.
14. Ferris, F.L., M.J. Podgor, and M.D. Davis, *Macular Edema in Diabetic Retinopathy Study Patients: Diabetic Retinopathy Study Report Number 12*. Ophthalmology, 1987. **94**(7): p. 754-760.
15. *Early Photocoagulation for Diabetic Retinopathy: ETDRS Report Number 9. Early Treatment Diabetic Retinopathy Study Research Group*. Ophthalmology, 1991. **98**(5, Supplement): p. 766-785.
16. *Photocoagulation Treatment of Proliferative Diabetic Retinopathy: Clinical Application of Diabetic Retinopathy Study (DRS) Findings, DRS Report Number 8. The Diabetic Retinopathy Study Research Group*. Ophthalmology, 1981. **88**(7): p. 583-600.
17. Thomas, R.L., et al., *IDF Diabetes Atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018*. Diabetes Res Clin Pract, 2019. **157**: p. 107840.
18. British Diabetic Association, *Retinal photography screening for diabetic eye disease*. London: BDA, 1997.
19. National Institute for Clinical Excellence, *Clinical Guideline E: Management of Type 2 diabetes*. 2002, NICE: London.

20. Teo, Z., et al., *Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis*. Ophthalmology, 2021. **128**.
21. Raman, R., et al., *Prevalence of diabetic retinopathy in India stratified by known and undiagnosed diabetes, urban-rural locations, and socioeconomic indices: results from the SMART India population-based cross-sectional screening study*. The Lancet Global Health, 2022. **10**(12): p. e1764-e1773.
22. Haider, S., et al., *Disease burden of diabetes, diabetic retinopathy and their future projections in the UK: cross-sectional analyses of a primary care database*. BMJ Open, 2021. **11**(7): p. e050058.
23. Liew, G., M. Michaelides, and C. Bunce, *A comparison of the causes of blindness certifications in England and Wales in working age adults (16–64 years), 1999–2000 with 2009–2010*. 2014. **4**(2): p. e004015.
24. Lee, R., T.Y. Wong, and C. Sabanayagam, *Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss*. Eye and Vision, 2015. **2**(1): p. 17.
25. Curran, K., et al., *Ophthalmologists' and patients' perspectives on treatments for diabetic retinopathy and maculopathy in Vietnam: a descriptive qualitative study*. BMJ Open, 2022. **12**(7): p. e055061.
26. Pearce, E. and S. Sivaprasad, *A Review of Advancements and Evidence Gaps in Diabetic Retinopathy Screening Models*. Clin Ophthalmol, 2020. **14**: p. 3285-3296.
27. Scanlon, P.H., *The English National Screening Programme for diabetic retinopathy 2003–2016*. Acta Diabetologica, 2017. **54**(6): p. 515-525.
28. Leese, G.P., et al., *Progression of Diabetes Retinal Status Within Community Screening Programs and Potential Implications for Screening Intervals*. Diabetes Care, 2014. **38**(3): p. 488-494.
29. Stratton, I.M., et al., *UKPDS 50: Risk factors for incidence and progression of retinopathy in Type II diabetes over 6 years from diagnosis*. Diabetologia, 2001. **44**(2): p. 156-163.
30. Frudd, K., et al., *Diagnostic circulating biomarkers to detect vision-threatening diabetic retinopathy: Potential screening tool of the future?* Acta Ophthalmol, 2022. **100**(3): p. e648-e668.
31. Aikaeli, F., et al., *Prevalence of microvascular and macrovascular complications of diabetes in newly diagnosed type 2 diabetes in low-and-middle-income countries: A systematic review and meta-analysis*. PLOS Global Public Health, 2022. **2**(6): p. e0000599.
32. Hoogeveen, E.K., *The Epidemiology of Diabetic Kidney Disease*. Kidney and Dialysis, 2022. **2**(3): p. 433-442.
33. De Boer, I.H., et al., *Temporal trends in the prevalence of diabetic kidney disease in the United States*. JAMA., 2011. **305**(24): p. 2532-2539.
34. Thomas, M.C., et al., *The burden of chronic kidney disease in Australian patients with type 2 diabetes (the NEFRON study)*. Med J Aust, 2006. **185**(3): p. 140-4.
35. Parving, H.H., et al., *Prevalence and risk factors for microalbuminuria in a referred cohort of type II diabetic patients: A global perspective*. Kidney International, 2006. **69**(11): p. 2057-2063.
36. Fiorentino, M., et al., *Renal biopsy in patients with diabetes: a pooled meta-analysis of 48 studies*. Nephrology Dialysis Transplantation, 2017. **32**(1): p. 97-110.
37. Afkarian, M., et al., *Kidney Disease and Increased Mortality Risk in Type 2 Diabetes*. Journal of the American Society of Nephrology, 2013. **24**(2).
38. Ding, Y. and M.E. Choi, *Autophagy in diabetic nephropathy*. The Journal of endocrinology, 2015. **224**(1): p. R15.
39. Mora-Fernández, C., et al., *Diabetic kidney disease: from physiology to therapeutics*. The Journal of physiology, 2014. **592**(18): p. 3997-4012.

40. Kyu, H.H., et al., *Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017*. The Lancet, 2018. **392**(10159): p. 1859-1922.
41. Levin, A., et al., *Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease*. Kidney international supplements, 2013. **3**(1): p. 1-150.
42. Yonata, A., et al., *Factors Affecting Quality of Life in Hemodialysis Patients*. Int J Gen Med, 2022. **15**: p. 7173-7178.
43. Kidney Disease: Improving Global Outcomes (KDIGO) Diabetes Work Group, *KDIGO 2020 Clinical Practice Guideline for Diabetes Management in Chronic Kidney Disease*. Kidney Int, 2020. **98**(4s): p. S1-s115.
44. Haider, S., et al., *Prognostic prediction models for diabetic retinopathy progression: a systematic review*. Eye (London, England), 2019. **33**(5): p. 702-713.
45. Gupta, M., et al., *Diabetic Retinopathy Is a Predictor of Progression of Diabetic Kidney Disease: A Systematic Review and Meta-Analysis*. Int J Nephrol, 2022. **2022**: p. 3922398.
46. Li, J., et al., *Correlations among Diabetic Microvascular Complications: A Systematic Review and Meta-analysis*. Sci Rep, 2019. **9**(1): p. 3137.
47. Yip, W., et al., *Retinal vascular imaging markers and incident chronic kidney disease: a prospective cohort study*. Scientific reports, 2017. **7**(1): p. 9374.
48. Yau, J.W.Y., et al., *Retinal arteriolar narrowing and subsequent development of CKD Stage 3: the Multi-Ethnic Study of Atherosclerosis (MESA)*. American Journal of Kidney Diseases, 2011. **58**(1): p. 39-46.
49. Lee, W., et al., *Ischemic diabetic retinopathy as a possible prognostic factor for chronic kidney disease progression*. Eye, 2014. **28**(9): p. 1119-1125.
50. Yau, J.W.Y., et al., *Retinal arteriolar narrowing and subsequent development of CKD Stage 3: the Multi-Ethnic Study of Atherosclerosis (MESA)*. American journal of kidney diseases : the official journal of the National Kidney Foundation, 2011. **58**(1): p. 39-46.
51. Zoungas, S., et al., *Effects of intensive glucose control on microvascular outcomes in patients with type 2 diabetes: a meta-analysis of individual participant data from randomised controlled trials*. The Lancet Diabetes & Endocrinology, 2017. **5**(6): p. 431-437.
52. Jin, J., et al., *Development of Diagnostic Biomarkers for Detecting Diabetic Retinopathy at Early Stages Using Quantitative Proteomics*. J Diabetes Res, 2016. **2016**: p. 6571976.
53. Kim, K., et al., *Verification of Biomarkers for Diabetic Retinopathy by Multiple Reaction Monitoring*. Journal of Proteome Research, 2010. **9**(2): p. 689-699.
54. Aspelund, T., et al., *Individual risk assessment and information technology to optimise screening frequency for diabetic retinopathy*. Diabetologia, 2011. **54**(10): p. 2525-32.
55. Sivaprasad, S., et al., *The ORNATE India Project: United Kingdom–India Research Collaboration to tackle visual impairment due to diabetic retinopathy*. Eye, 2020. **34**(7): p. 1279-1286.
56. Sivaprasad, S., et al., *Protocol on a multicentre statistical and economic modelling study of risk-based stratified and personalised screening for diabetes and its complications in India (SMART India)*. 2020. **10**(12): p. e039657.
57. Gurudas, S., et al., *National prevalence of vision impairment and blindness and associated risk factors in adults aged 40 years and older with known or undiagnosed diabetes: results from the SMART-India cross-sectional study*. The Lancet Global Health, 2024.
58. Khan, R., et al., *Prevalence and incidence of visual impairment in patients with proliferative diabetic retinopathy in India*. Sci Rep, 2020. **10**(1): p. 10513.

59. Nugawela, M.D., et al., *Ethnic Disparities in the Development of Sight-Threatening Diabetic Retinopathy in a UK Multi-Ethnic Population with Diabetes: An Observational Cohort Study*. J Pers Med, 2021. **11**(8).
60. Blighe, K., et al., *Diabetic Retinopathy Environment-Wide Association Study (EWAS) in NHANES 2005-2008*. J Clin Med, 2020. **9**(11).
61. Gurudas, S., et al., *Multicenter Evaluation of Diagnostic Circulating Biomarkers to Detect Sight-Threatening Diabetic Retinopathy*. JAMA Ophthalmol, 2022. **140**(6): p. 587-597.
62. Nugawela, M.D., et al., *Development and validation of predictive risk models for sight threatening diabetic retinopathy in patients with type 2 diabetes to be applied as triage tools in resource limited settings*. EClinicalMedicine, 2022. **51**: p. 101578.
63. Gurudas, S., et al., *Development and validation of resource-driven risk prediction models for incident chronic kidney disease in type 2 diabetes*. Sci Rep, 2021. **11**(1): p. 13654.
64. Yau, J.W., et al., *Global prevalence and major risk factors of diabetic retinopathy*. Diabetes Care, 2012. **35**(3): p. 556-64.
65. Wilkinson, C.P., et al., *Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales*. Ophthalmology, 2003. **110**(9): p. 1677-1682.
66. Vashist, P., et al., *Blindness and visual impairment and their causes in India: Results of a nationally representative survey*. PLOS ONE, 2022. **17**(7): p. e0271736.
67. Purola, P., S. Koskinen, and H. Uusitalo, *Impact of vision on generic health-related quality of life – A systematic review*. Acta Ophthalmol.
68. Tandon, N., et al., *The increasing burden of diabetes and variations among the states of India: the Global Burden of Disease Study 1990-2016*. The Lancet Global Health, 2018. **6**(12): p. e1352-e1362.
69. Sivaprasad, S., et al., *Protocol on a multicentre statistical and economic modelling study of risk-based stratified and personalised screening for diabetes and its complications in India (SMART India)*. BMJ Open, 2020. **10**(12): p. e039657.
70. Lumley, T., *Analysis of Complex Survey Samples*. Journal of Statistical Software, 2004. **9**(1): p. 1-19.
71. Binder, D.A., *On the Variances of Asymptotically Normal Estimators from Complex Surveys*. International Statistical Review / Revue Internationale de Statistique, 1983. **51**(3): p. 279-292.
72. Hosmer, D.W. and S. Lemeshow, *Applied Logistic Regression*. 2004: Wiley.
73. GBD 2019 Blindness and Vision Impairment Collaborators; Vision Loss Expert Group of the Global Burden of Disease Study, *Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study*. Lancet Glob Health, 2021. **9**(2): p. e144-e160.
74. Vashist, P., et al., *Prevalence of diabetic retinopathy in India: Results from the National Survey 2015-19*. Indian Journal of Ophthalmology, 2021. **69**(11): p. 3087.
75. Barrett, J.S. and P.M. Heaton, *Real-World Data: An Unrealized Opportunity in Global Health?* Clin Pharmacol Ther, 2019. **106**(1): p. 57-59.
76. India, O.o.R.G.a.C.C. *Population finder*. 2011; Available from: <https://censusindia.gov.in/census.website/data/population-finder>.
77. Rothman, K. and S. Greenland, *Modern Epidemiology* 1986, Little, Brown: Boston, MA.
78. Scanlon, P.H., S.J. Aldington, and I.M. Stratton, *Delay in diabetic retinopathy screening increases the rate of detection of referable diabetic retinopathy*. Diabet Med, 2014. **31**(4): p. 439-42.
79. Rani, P.K., et al., *Prevalence of Visual Impairment and Associated Risk Factors in Subjects with Type II Diabetes Mellitus: Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study (SN-DREAMS, Report 16)*. Middle East Afr J Ophthalmol, 2012. **19**(1): p. 129-34.

80. Lopez-Ramos, A., et al., *Rapid assessment of avoidable blindness: Prevalence of blindness, visual impairment and diabetes in nuevo leon, Mexico 2014*. Ophthalmic Epidemiol, 2018. **25**(5-6): p. 412-418.
81. Bhatt, B.R., *ASHAs in rural India, the ray of hope for diabetes care*. Journal of Social Health and Diabetes, 2014. **2**: p. 18-24.
82. Chokshi, M., et al., *Health systems in India*. J Perinatol, 2016. **36**(s3): p. S9-s12.
83. Kansora, M.B. and R. Goldhardt, *Decision Making in Proliferative Diabetic Retinopathy Treatment*. Curr Ophthalmol Rep, 2019. **7**(1): p. 45-50.
84. Scanlon, P.H., et al., *The influence of age, duration of diabetes, cataract, and pupil size on image quality in digital photographic retinal screening*. Diabetes Care, 2005. **28**(10): p. 2448-53.
85. Conlin, P., et al., *Non-mydratic teleretinal imaging improves adherence to annual eye examinations in patients with diabetes*. Journal of rehabilitation research and development, 2006. **43**: p. 733-40.
86. Ornate India. *Moorfields project changes public health policy in India*. 2019 08/08/2023]; Available from: <https://ornateindia.net/international-meet-discusses-the-issue-of-vision-loss-caused-due-to-diabetes/>.
87. Lartey, S.Y. and A.K. Aikins, *Visual impairment amongst adult diabetics attending a tertiary outpatient clinic*. Ghana Med J, 2018. **52**(2): p. 84-87.
88. Choovuthayakorn, J., et al., *Characteristics and Outcomes of Pars Plana Vitrectomy for Proliferative Diabetic Retinopathy Patients in a Limited Resource Tertiary Center over an Eight-Year Period*. J Ophthalmol, 2019. **2019**: p. 9481902.
89. Hull, S.A., et al., *Recording ethnicity in primary care: assessing the methods and impact*. Br J Gen Pract, 2011. **61**(586): p. e290-4.
90. Von Elm, E., et al., *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies*. The Lancet, 2007. **370**(9596): p. 1453-1457.
91. Weiskopf, N.G., et al., *A Data Quality Assessment Guideline for Electronic Health Record Data Reuse*. EGEMS (Wash DC), 2017. **5**(1): p. 14.
92. Kalbfleisch, J.D. and D.E. Schaubel, *Fifty Years of the Cox Model*. Annual Review of Statistics and Its Application, 2023. **10**(1): p. 1-23.
93. Cox, D.R., *Regression models and life-tables*. Journal of the Royal Statistical Society: Series B, 1972. **34**(2): p. 187-202.
94. Nardi, A. and M. Schemper, *Comparing Cox and parametric models in clinical studies*. Stat Med, 2003. **22**(23): p. 3597-610.
95. Mathur, R., et al., *Population trends in the 10-year incidence and prevalence of diabetic retinopathy in the UK: a cohort study in the Clinical Practice Research Datalink 2004–2014*. BMJ Open, 2017. **7**(2): p. e014444.
96. Sivaprasad, S., et al., *Ethnic variations in the prevalence of diabetic retinopathy in people with diabetes attending screening in the United Kingdom (DRIVE UK)*. PLoS One, 2012. **7**(3): p. e32182.
97. Scobie, S., J. Spencer, and V.J.N.T. Raleigh, *Ethnicity coding in English health service datasets*. Research report, Nuffield Trust, 2021.
98. National Health Service (NHS), *The National Health Service (General Medical Services Contracts and Personal Medical Services Agreements) (Amendment) (No. 3) Regulations 2020*. 2020.
99. Bhatwadekar, A.D., et al., *Genetics of Diabetic Retinopathy, a Leading Cause of Irreversible Blindness in the Industrialized World*. Genes (Basel), 2021. **12**(8).
100. Homer, K., et al., *Making ends meet - relating a self-reported indicator of financial hardship to health status*. Journal of public health (Oxford, England), 2023. **45**.

101. Adams, J., V. Ryan, and M. White, *How accurate are Townsend Deprivation Scores as predictors of self-reported health? A comparison with individual level data*. J Public Health (Oxf), 2005. **27**(1): p. 101-6.
102. Pardo-Crespo, M.R., et al., *Comparison of individual-level versus area-level socioeconomic measures in assessing health outcomes of children in Olmsted County, Minnesota*. J Epidemiol Community Health, 2013. **67**(4): p. 305-10.
103. Stafford, M. and M. Marmot, *Neighbourhood deprivation and health: does it affect us all equally?* International Journal of Epidemiology, 2003. **32**(3): p. 357-366.
104. Sivaprasad, S., et al., *Prevalence of diabetic retinopathy in various ethnic groups: a worldwide perspective*. Surv Ophthalmol, 2012. **57**(4): p. 347-70.
105. Efron, B., *The Efficiency of Cox's Likelihood Function for Censored Data*. Journal of the American Statistical Association, 1977. **72**(359): p. 557-565.
106. Oakes, D., *The asymptotic information in censored survival data*. Biometrika, 1977. **64**(3): p. 441-448.
107. Cox, D.R. and D. Oakes, *Analysis of survival data*. Vol. 21. 1984: CRC press.
108. Turnbull, B.W., *The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1976. **38**(3): p. 290-295.
109. StataCorp, *Stata Statistical Software: Release 17*. 2021, StataCorp LLC: College Station, TX.
110. Chen, X., et al., *Serum uric acid concentration is associated with hypertensive retinopathy in hypertensive chinese adults*. BMC Ophthalmol, 2017. **17**(1): p. 83.
111. Luo, B.A., F. Gao, and L.L. Qin, *The Association between Vitamin D Deficiency and Diabetic Retinopathy in Type 2 Diabetes: A Meta-Analysis of Observational Studies*. Nutrients, 2017. **9**(3).
112. Kong, X., et al., *Association between Free Thyroxine Levels and Diabetic Retinopathy in Euthyroid Patients with Type 2 Diabetes Mellitus*. Endocr Res, 2020. **45**(2): p. 111-118.
113. Merin, S. and M. Freund, *Retinopathy in severe anemia*. Am J Ophthalmol, 1968. **66**(6): p. 1102-6.
114. Khan, A.A., A.H. Rahmani, and Y.H. Aldebasi, *Diabetic Retinopathy: Recent Updates on Different Biomarkers and Some Therapeutic Agents*. Curr Diabetes Rev, 2018. **14**(6): p. 523-533.
115. Chatziralli, I.P., *The Role of Dyslipidemia Control in the Progression of Diabetic Retinopathy in Patients with Type 2 Diabetes Mellitus*. Diabetes Ther, 2017. **8**(2): p. 209-212.
116. *Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33)*. UK Prospective Diabetes Study (UKPDS) Group. The Lancet, 1998. **352**(9131): p. 837-853.
117. Kim, K., et al., *Verification of multimarkers for detection of early stage diabetic retinopathy using multiple reaction monitoring*. J Proteome Res, 2013. **12**(3): p. 1078-89.
118. Patel, C.J., J. Bhattacharya, and A.J. Butte, *An environment-wide association study (EWAS) on type 2 diabetes mellitus*. J PloS one, 2010. **5**(5): p. e10746.
119. Centers for Disease Control and Prevention, *National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire (or Examination Protocol, or Laboratory Protocol)*. 2006.
120. Zhuang, X., et al., *Toward a panoramic perspective of the association between environmental factors and cardiovascular disease: An environment-wide association study from National Health and Nutrition Examination Survey 1999-2014*. Environ Int, 2018. **118**: p. 146-153.

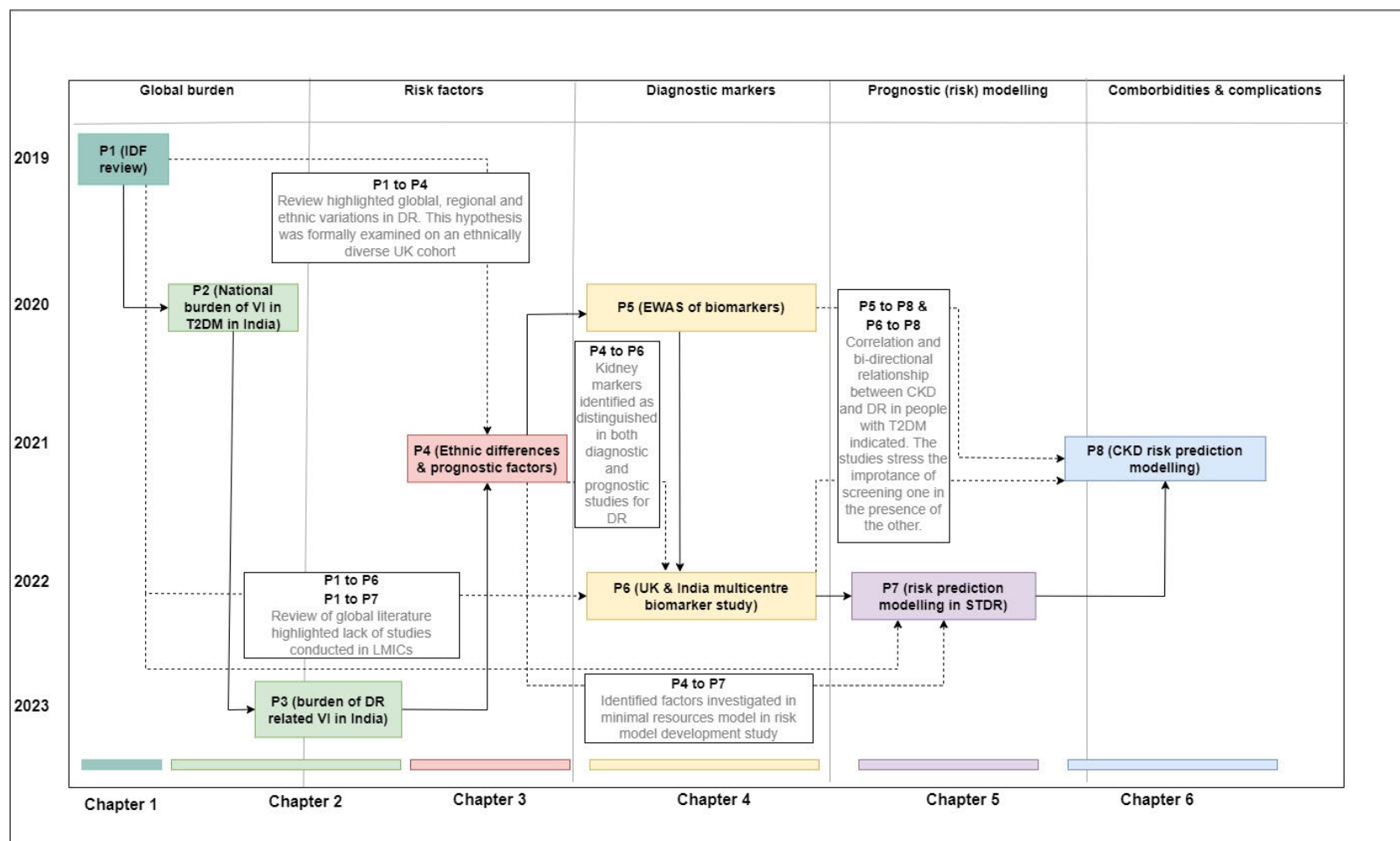
121. Engel, J., L. Buydens, and L. Blanchet, *An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics*. Journal of Chemometrics, 2017. **31**(4): p. e2880.
122. Zou, H. and T. Hastie, *Regularization and Variable Selection Via the Elastic Net*. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005. **67**(2): p. 301-320.
123. Heydon, P., et al., *Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients*. Br J Ophthalmol, 2021. **105**(5): p. 723-728.
124. Devi, K.S.S. and U.M. Krishnan, *Microfluidic electrochemical immunosensor for the determination of cystatin C in human serum*. Mikrochim Acta, 2020. **187**(10): p. 585.
125. Bzdok, D., *Classical Statistics and Statistical Learning in Imaging Neuroscience*. Front Neurosci, 2017. **11**: p. 543.
126. Bzdok, D., M. Krzywinski, and N. Altman, *Points of Significance: Machine learning: a primer*. Nat Methods, 2017. **14**(12): p. 1119-1120.
127. Zheng, Y., et al., *Design and methodology challenges of environment-wide association studies: A systematic review*. Environ Res, 2020. **183**: p. 109275.
128. Mighty, J., et al., *Extracellular vesicles of human diabetic retinopathy retinal tissue and urine of diabetic retinopathy patients are enriched for the junction plakoglobin protein*. Front Endocrinol (Lausanne), 2022. **13**: p. 1077644.
129. JAMA Ophthalmology, *JN Learning*, in *Interview with Sobha Sivaprasad, DM, author of Multicenter Evaluation of Diagnostic Circulating Biomarkers to Detect Sight-Threatening Diabetic Retinopathy*. Hosted by Neil Bressler, MD., N. Bressler and S. Sivaprasad, Editors. 2022.
130. Ting, D.S., G.C. Cheung, and T.Y. Wong, *Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review*. Clin Exp Ophthalmol, 2016. **44**(4): p. 260-77.
131. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement*. BMC Medicine, 2015. **13**(1): p. 1.
132. Steyerberg, E.W., et al., *Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research*. PLOS Medicine, 2013. **10**(2): p. e1001381.
133. Zeng, D., L. Mao, and D. Lin, *Maximum Likelihood Estimation for Semiparametric Transformation Models with Interval-Censored Data*. Biometrika, 2016. **103**.
134. StataCorp, *Stata 18 Survival Analysis Reference Manual*. . 2023, Stata Press: College Station, TX.
135. Zeng, D., L. Mao, and D.Y. Lin, *Maximum likelihood estimation for semiparametric transformation models with interval-censored data*. Biometrika, 2016. **103**(2): p. 253-271.
136. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation*. Eur Heart J, 2014. **35**(29): p. 1925-31.
137. Hippisley-Cox, J., C. Coupland, and P. Brindle, *Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study*. BMJ, 2017. **357**: p. j2099.
138. George, C., et al., *The need for screening, early diagnosis, and prediction of chronic kidney disease in people with diabetes in low- and middle-income countries—a review of the current literature*. BMC Medicine, 2022. **20**: p. 247.
139. Royston, P., G. Ambler, and W. Sauerbrei, *The use of fractional polynomials to model continuous risk variables in epidemiology*. International Journal of Epidemiology, 1999. **28**(5): p. 964-974.



140. Royston, P. and D.G. Altman, *Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling*. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1994. **43**(3): p. 429-467.
141. Royston, P. and G. Ambler, *Multivariable fractional polynomials*. Stata Technical Bulletin, 1999. **8**(43).
142. Vickers, A.J. and E.B. Elkin, *Decision curve analysis: a novel method for evaluating prediction models*. J Medical Decision Making, 2006. **26**(6): p. 565-574.
143. Vickers, A.J., B. Van Calster, and E.W. Steyerberg, *Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests*. BMJ, 2016. **352**.
144. Bonnett, L.J., et al., *Guide to presenting clinical prediction models for use in clinical settings*. BMJ, 2019. **365**: p. l737.
145. Cabrera, A.P., et al., *Do Genomic Factors Play a Role in Diabetic Retinopathy?* J Clin Med, 2020. **9**(1).
146. Department of Health, *NHS Reference Costs 2014-2015*, D.o.H.a.S. Care, Editor. 2015.
147. National Health Service (NHS), *Delivery plan for tackling the COVID-19 backlog of elective care*. 2022: England.
148. National Institute for Health and Care Excellence (NICE), *Type 2 diabetes in adults: management [NICE Guideline No. 28]*. 2015, ammended 2022.
149. Thomas, R.L., et al., *Cost-effectiveness of biennial screening for diabetes related retinopathy in people with type 1 and type 2 diabetes compared to annual screening*. Eur J Health Econ, 2020. **21**(7): p. 993-1002.
150. Stratton, I.M., et al., *A simple risk stratification for time to development of sight-threatening diabetic retinopathy*. Diabetes Care, 2013. **36**(3): p. 580-5.
151. Lund, S.H., et al., *Individualised risk assessment for diabetic retinopathy and optimisation of screening intervals: a scientific approach to reducing healthcare costs*. Br J Ophthalmol, 2016. **100**(5): p. 683-7.
152. James, M., et al., *Cost effectiveness analysis of screening for sight threatening diabetic eye disease*. BMJ, 2000. **320**(7250): p. 1627-31.
153. Liew, G., M. Michaelides, and C. Bunce, *A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999-2000 with 2009-2010*. BMJ Open, 2014. **4**(2): p. e004015.
154. Haleem, A., et al., *Biosensors applications in medical field: A brief review*. Sensors International, 2021. **2**: p. 100100.
155. Collins, G.S., et al., *External validation of multivariable prediction models: a systematic review of methodological conduct and reporting*. BMC Med Res Methodol, 2014. **14**: p. 40.
156. PramodKumar, T.A., et al., *Role of cystatin C in the detection of sight-threatening diabetic retinopathy in Asian Indians with type 2 diabetes*. Journal of Diabetes and its Complications, 2023. **37**(8): p. 108545.
157. Sciutto, G., et al., *Miniaturized Biosensors to Preserve and Monitor Cultural Heritage: from Medical to Conservation Diagnosis*. Angew Chem Int Ed Engl, 2018. **57**(25): p. 7385-7389.
158. Kim, J., et al., *Wearable biosensors for healthcare monitoring*. Nat Biotechnol, 2019. **37**(4): p. 389-406.
159. Lip, G.Y., et al., *Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation*. Chest, 2010. **137**(2): p. 263-72.
160. Sharma, V., et al., *Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records*. BMJ Health Care Inform, 2021. **28**(1): p. e100253.

161. Hippisley-Cox, J., et al., *Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2*. BMJ, 2008. **336**(7659): p. 1475-1482.
162. Hook, D.W., S.J. Porter, and C. Herzog, *Dimensions: Building Context for Search and Evaluation*. Frontiers in Research Metrics and Analytics, 2018. **3**: p. 23.
163. Digital Science. *Altmetric*. 12/11/2023]; Available from: <https://www.altmetric.com>.
164. Plum Analytics. *Plumx Metrics*. 12/11/2023]; Available from: <http://plumanalytics.com/products/plumx-metrics/>.
165. Coventry University. *Pure Portal*. 12/11/2023]; Available from: <https://pureportal.coventry.ac.uk/en/persons/sarega-gurudas/publications/>.

## Appendix 1. Publications flow diagram showing relationships between manuscripts



Abbreviations: IDF- International Diabetes Federation, DR- Diabetic Retinopathy, T2DM- Type 2 Diabetes Mellitus, EWAS- Environment Wide Association Study, LMIC- Low- and Middle-Income Country, STDR- Sight Threatening Diabetic Retinopathy, CKD- Chronic Kidney Disease

## Appendix 2. STARD guidelines for Publication 6

Section & Topic	No	Item	Reported on page #
<b>TITLE OR ABSTRACT</b>			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1 - ROC
<b>ABSTRACT</b>			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	1
<b>INTRODUCTION</b>			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	2 – Biomarkers to be used as a potential triage tool for retinal screening
	4	Study objectives and hypotheses	2 – To evaluate previously verified blood-based biomarkers using ELISA for their usefulness as indicators for STDR
<b>METHODS</b>			
<i>Study design</i>	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	2- Prospective
<i>Participants</i>	6	Eligibility criteria	5 – Outcome groups
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	2 – Outpatient ophthalmology clinics based on diagnosis in electronic health records. Normal participants without

			diabetes may include patients and non-patients (bystanders).
	<b>8</b>	Where and when potentially eligible participants were identified (setting, location and dates)	2- Outpatient ophthalmology clinics
	<b>9</b>	Whether participants formed a consecutive, random or convenience series	Convenience series
<i>Test methods</i>	<b>10a</b>	Index test, in sufficient detail to allow replication	2- Blood biomarkers, laboratory tests and potential confounders
	<b>10b</b>	Reference standard, in sufficient detail to allow replication	2- Sight threatening diabetic retinopathy vs no diabetic retinopathy
	<b>11</b>	Rationale for choosing the reference standard (if alternatives exist)	2- Biomarkers assessed based on prior literature review. Previous biomarker studies have been on small sample sizes (~60 patients)
	<b>12a</b>	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	5-6 - Exploratory cut-offs based on achieving sensitivity and specificity of desired values (80%, 90%) and Youden index
	<b>12b</b>	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	5-Diabetic retinopathy grading (ETDRS)

	<b>13a</b>	Whether clinical information and reference standard results were available to the performers/readers of the index test	2-Laboratory staff were masked to the clinical diagnosis and data
	<b>13b</b>	Whether clinical information and index test results were available to the assessors of the reference standard	Index test results were carried out after reference standard assessments
<i>Analysis</i>	<b>14</b>	Methods for estimating or comparing measures of diagnostic accuracy	6 – AUC , sensitivity, specificity, PPV, NPV, LR+, LR-
	<b>15</b>	How indeterminate index test or reference standard results were handled	None
	<b>16</b>	How missing data on the index test and reference standard were handled	Missing data were handled using complete-case analysis
	<b>17</b>	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	6 – sensitivity analyses comparing diagnostic accuracy of Cystatin C in those with normal kidney function to those with kidney impairment
	<b>18</b>	Intended sample size and how it was determined	5
<b>RESULTS</b>			
<i>Participants</i>	<b>19</b>	Flow of participants, using a diagram	eFigure 1
	<b>20</b>	Baseline demographic and clinical characteristics of participants	6, Table 1
	<b>21a</b>	Distribution of severity of disease in those with the target condition	6, Table 1
	<b>21b</b>	Distribution of alternative diagnoses in those without the target condition	6, Table 1
	<b>22</b>	Time interval and any clinical interventions between index test and reference standard	None

<i>Test results</i>	<b>23</b>	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	6-8, Table 1, 2
	<b>24</b>	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	8-9, Table 2, 3
	<b>25</b>	Any adverse events from performing the index test or the reference standard	None
<b>DISCUSSION</b>			
	<b>26</b>	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	11
	<b>27</b>	Implications for practice, including the intended use and clinical role of the index test	11- Triage test (pre-screening model) in poorly resourced settings
<b>OTHER INFORMATION</b>			
	<b>28</b>	Registration number and name of registry	
	<b>29</b>	Where the full study protocol can be accessed	Not published
	<b>30</b>	Sources of funding and other support; role of funders	Grant MR/P027881/1

### Appendix 3. 7 steps to clinical risk prediction development and validation

Stage	Method	
	STDR Prognostic model (P7)	CKD prognostic model (P8)
i) <i>consideration of the research question and initial data inspection</i>	<ul style="list-style-type: none"> <li>- Gathered evidence on the limitations of current STDR prediction models for use in resource-restricted settings</li> </ul>	<ul style="list-style-type: none"> <li>- Gathered evidence on the limitations of current CKD prediction models for use in resource-restricted settings</li> </ul>
ii) <i>coding of predictors</i>	<ul style="list-style-type: none"> <li>- Non-linear predictors in <b>P7</b> were categorised based on clinically defined cut-points for it to provide risk estimates based on categories so that it can be used in risk charts. As the models used minimal number of variables, translating the models into a risk chart for use in community screening meant that the variables needed to be presented in categories.</li> <li>- Age was categorised into clinically relevant risk groups, in part due to increased sample size in <b>P7</b> which allowed each age group to be modelled with sufficient events per variable and due to interaction between age and duration, there was less chance of model misspecification if both terms were not modelled as continuous</li> </ul>	<ul style="list-style-type: none"> <li>- Non-linear predictors in <b>P8</b> were modelled as continuous variables instead of categorisation based on clinically defined cut-points. This allowed the models to retain good model accuracy, where categorizing may often lead to loss of information, and consequently lower model c-statistic.</li> <li>- eGFR was categorized as in EHR data values were truncated at 90 ml/min/1.73m<sup>2</sup>, therefore modelling the variable as continuous would lead to biased estimates.</li> <li>- ACR was categorized using thresholds used in CKD guidelines, to limit bias due to outliers. Interaction between age and insulin-use was modelled due to improvement in model c-statistic from its inclusion.</li> </ul>
iii) <i>model specification</i>		



iv) <i>model estimation</i>	<ul style="list-style-type: none"> <li>- Cox proportional hazards model was used</li> <li>- Does not parametrize or assume a distribution for the baseline hazard</li> <li>- A robust and widely used modelling technique for survival data</li> <li>- Sensitivity analysis using interval-censored cox models to model interval-censored nature of routine healthcare data in PwD</li> </ul>	<ul style="list-style-type: none"> <li>- Cox proportional hazards model was used.</li> <li>- Does not parametrize or assume a distribution for the baseline hazard. A robust and widely used modelling technique for survival data.</li> </ul>
v) <i>evaluation of model performance</i>	<ul style="list-style-type: none"> <li>- Concordance-statistic (discrimination), observed to expected ratio (calibration), beta-coefficient of the calibration slope (calibration)</li> </ul>	Concordance-statistic (discrimination), observed to expected ratio measures (calibration), beta-coefficient of the calibration slope (calibration)
vi) <i>model validation</i>	<ul style="list-style-type: none"> <li>- External validation assessed in two independent cohorts (MDRF dataset and SAIL databank)</li> </ul>	<ul style="list-style-type: none"> <li>- Optimism adjusted for using 10-fold cross-validation of Harrell's concordance-statistic (c-statistic) and calibration slope, to minimise the risk of overfitting bias. External validation in the SAIL databank.</li> </ul>
vii) <i>model presentation</i>	<ul style="list-style-type: none"> <li>- Risk models supplemented with a risk chart as a visual aid for community workers</li> </ul>	Converted risk models into a points-based risk score for ease of interpretation.

Modified from Steyerberg et al [136]. Abbreviations: SAIL- Secure Anonymised Information Linkage, MDRF-Madras Diabetes Research Foundation, eGFR- estimated Glomerular Filtration Rate, ACR- Albumin to creatinine ratio, EHR- Electronic Health Record, CKD- Chronic Kidney Disease