Mobile health for cardiovascular disease risk prediction and management in resource-constrained environments

Arvind Raghu

Department of Engineering Science

University of Oxford

Supervised by
Prof. Lionel Tarassenko, Prof. Gari Clifford

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfilment of the requirements for the degree of

Doctor of Philosophy

Worcester College

Michaelmas Term, 2015

Arvind Raghu
Worcester College
Doctor of Philosophy
Michaelmas Term, 2015

Mobile health for cardiovascular disease risk prediction and management in resource-constrained environments

Abstract

It is well established that the leading global cause of mortality and morbidity, cardiovascular disease (CVD), is more severe in resource-constrained environments such as rural India (RI). This thesis explores how best to manage CVD risk in RI by using a mobile-based, point-of-care (POC) Clinical Decision Support System (CDSS), *SMARThealth*, that is designed to assist Accredited Social Health Activists (ASHAs) or minimally trained health workers. The four major focus areas are:

(a) Design, development, and large-scale data collection using *SMARThealth* - an agile development process and user-centred design approach were followed to pilot test the CDSS with 292 participants. Evaluation metrics included system efficiency, end-user variability, usability, and sub-group analysis to identify better or poorly performing ASHAs. An improved version of *SMARThealth* was used for baseline data collection across 54 villages (62,194 participants) in Andhra Pradesh, India. 9864 (15.8%) of the participants were at high CVD risk.

(b) Improvement of the sole CVD risk prediction algorithm for RI, the WHO/ISH CVD risk prediction charts (WHO-ISHc) - the choice of the low information (LI) model or high information (HI) model of WHO-ISHc was statistically significant for CVD risk prediction in RI ($p=0.008; \chi^2=7.03$) with 155 subjects (or 14.5% of 1066 patients) having different CVD risk scores according to the LI and HI WHO-ISHc. A parsimonious POC test was developed to identify patients for whom risk prediction by the HI and LI WHO/ISHc differ (that is, for whom the assessment of total cholesterol would be beneficial). The POC test showed good discrimination (out-of-sample AUC 0.85 with Random Forests).

(c) Assessment of best prediction algorithm for RI - eight highly predictive features of CVD risk were identified based on labelled data, and the resulting model (*Model 1*) had higher or equal AUCs and log-likelihood scores, and lower Brier scores when compared to a benchmark algorithm. The contribution of age and gender alone offered good discrimination and recalibration of *Model 1* for RI was introduced. The lack of recorded end outcomes in RI prompted the use of an unsupervised approach to identify high-risk patients. Clusters of low and high CVD risks were found when $\hat{K}=2$, but also clusters with intermediate risk when $\hat{K}=4$ offering an alternative approach to identifying groups of high-risk patients.

(d) Analysis from a randomised controlled trial evaluation of *SMARThealth* - preliminary data analysis of 131 high-risk patients during the first year of the randomised controlled trial showed a statistically significant reduction in median blood pressure between the 1^{st} and 5^{th} assessment (p=0.0097). The proportion of patients under treatment for high blood pressure continued to increase throughout.

Related Publications

Journal Publications

1. Raghu, Arvind, Devarsetty Praveen, David Peiris, Lionel Tarassenko, and Gari Clifford. "Implications of Cardiovascular Disease Risk Assessment Using the WHO/ISH Risk Prediction Charts in Rural India". *PLoS ONE* 10, no.8 (2015)

2. Raghu, Arvind, Devarsetty Praveen, David Peiris, Lionel Tarassenko, and Gari Clifford. "Engineering a mobile health tool for resource-poor settings to assess and manage cardiovascular disease risk: SMARThealth study." *BMC Medical Informatics and Decision Making* 15, no.1 (2015).

3. Praveen, Devarsetty, Anushka Patel, Arvind Raghu, Gari D. Clifford, Pallab K. Maulik, Ameer Mohammad Abdul, Kishor Mogulluru, Lionel Tarassenko, Stephen MacMahon, and David Peiris. "SMARTHealth India: Development and Field Evaluation of a Mobile Clinical Decision Support System for Cardiovascular Diseases in Rural India." *Journal of Medical Internet Research mHealth and uHealth* 2, no.4 (2014).

Book chapters

1. Raghu, Arvind, Devarsetty Praveen, David Peiris, Lionel Tarassenko, and Gari Clifford. "Lessons from the evaluation of a clinical decision support tool for cardiovascular disease risk management in rural India." In *Technologies for Development*, pp. 199-209. Springer International Publishing, Switzerland (2015).

2. Raghu, Arvind and Gari Clifford. "Role of ICT in community mobilization and health promotion" In Global Health Informatics, *MIT Press*, Cambridge, MA, USA (2015) *In Press*

Conference papers

1. Raghu, Arvind, Devarsetty Praveen, David Peiris, Lionel Tarassenko, and Gari Clifford. “Mobile health for Cardiovascular disease risk screening and management in resource-constrained environments." *Annual Conference on Mobile and Information Technologies in Medicine and Health*, Prague, Czech Republic (2013)

Contents

List of Figures

List of Tables

Nomenclature

List of frequently used acronyms / abbreviations

APHRI The Andhra Pradesh Rural Health Initiative[3], a cross-sectional study of CVD risk-factors in rural Andhra Pradesh

ASHA Accredited Social Health Activist

ASSIGN ASsessing cardiovascular risk using Scottish Intercollegiate Guidelines Network (SIGN) guidelines

AUC Area Under Curve

BMI Body Mass Index

BP Blood Pressure

CDSS Clinical Decision Support System

CVD Cardiovascular disease

DBP Diastolic Blood Pressure

FRS Framingham Risk Score (sometimes 1,2, or 3 is suffixed to denote versions of the FRS)

GDP Gross Domestic Product

HDL High Density Lipoprotein cholesterol

HI High Information

ICT Information and Communications Technology

IQR InterQuartile Range

ISH International Society for Hypertension

LDL Low Density Lipoprotein cholesterol

LI Low Information

LMIC Lower and Middle Income Countries

LR Logistic Regression

mHealth mobile health

mRMR Maximum Relevance Minimum Redundancy feature selection technique

NPCDCS The Indian National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular diseases and Stroke

PHC Primary Health Centres

POC Point-Of-Care

PROCAM Prospective CArdiovascular Munster score

PVD Peripheral Vascular Disease

QRISK QRESEARCH cardiovascular Risk. The QRESEARCH database is a large consolidated collection of health records from UK general practitioners

RCT Randomized Controlled Trial

RF Random Forest

SBP Systolic Blood Pressure

SCORE Systematic COronary Risk Evaluation

SEAR − D South East Asian Regions D which includes Bangladesh, Bhutan, Democratic People's Republic of Korea, India, Maldives, Myanmar, and Nepal

SMARThealth Systematic Medical Assessment, Referral, and Treatment health programme

SVM Support Vector Machine

swc − RCT Stepped Wedge Cluster Randomized Controlled Trial

TC Total Cholesterol

TG TriGlycerides

WHO World Health Organization

Chapter 1

Motivation and Background

1.1 Global burden of Cardiovascular disease (CVD)

The leading cause of death and disability worldwide is Cardiovascular disease (CVD) [4]. CVD is a broad term, covering any disorder related to the heart and the circulatory system. This can refer to coronary artery disease such as myocardial infarction or angina, cerebrovascular disease such as stroke or transient ischaemic attacks, or peripheral vascular disease. Lower and Middle Income Countries (LMICs) share the majority of the CVD burden with over 80% of CVD deaths occuring in these countries [4]. At the same time, the amount of money spent on healthcare as a percentage of GDP is substantially lower in LMICs ($72 per capita) as compared to developed countries ($2700 per capita) [5]. An example of this is the fifty-fold difference between the USA and South Africa with regard to spending on CVD care [5]. A majority of CVD events are preventable, yet current prevention methods are inadequate to substantially reduce the global burden [4].

1.2 CVD in India

Despite the country being host to 17% of the world's population, India's public spending on healthcare is less than 1% of the world's total health expenditure [6]. Statistics show that between 80-86% of health-related expenditure in India is met out-of-pocket [7][8], one of the highest in the world. Furthermore, a majority of the Indian public resides in rural areas where there is often a disproportionate ratio of doctors to patients. The acute shortage of primary-care physicians leads to many hurdles in healthcare delivery, especially in rural areas. It is estimated that out of 1.1 million medical practitioners in India, about 75% of them live in urban areas and 700 Primary Health Centres (PHCs) are without a physician [9]. This clear urban-rural, rich-poor divide greatly impairs the ability of many sectors of the population to access quality healthcare, such as the urban-poor, who find it extremely difficult to meet the costs of private health care flourishing in the cities. According to the Indian government's Insurance Regulatory and Development Authority, only 216.2 million or 17% of the population was covered by health insurance as of March 2014 [10].

The burden of CVD in India is profound. The annual number of years of life lost due to coronary heart disease related-deaths before the age of 60 years was 7·1 million in 2004 and is estimated to increase to 17·9 million in 2030 in India [11]. This number is greater than the sum of projected life years lost due to the same cause in China, Russia, and the USA combined [11]. Hypertension, a major risk factor for CVD, is responsible for 16% of ischaemic heart disease, 21% of peripheral vascular disease, 24% of acute myocardial infarctions and 29% of strokes in India according to estimates from the Indian Council of Medical Research [12]. The burden of hypertension in India is such that the estimated number of people with hypertension (118 million diagnosed in 2000) will double by 2025 (projected to be 212 million) [13]. Furthermore, by 2025, it is estimated that 189 million

Indians will be over 60 years of age (triple the number in 2004), placing an additional burden on the already strained healthcare infrastructure [14]. Also there has been a shift in attention from communicable diseases to chronic diseases as the rise in lifestyle related illnesses such as stroke, heart disease, hyperlipidemia, diabetes and hypertension reaches epidemic proportions.

1.3 Primary care provisions for managing CVD in rural India

Since independence in 1947, there have been various health committee reports in India as well as five-year plans to address the country's healthcare needs. In 1978, India was a signatory of the Alma Ata declaration which carried the motto "Health for all by year 2000". This prompted the formulation of India's first National Health Policy (NHP) in 1983 which attempted to incorporate a universal and comprehensible primary health care system. The second NHP in 2002 aimed to be a progression of the first. However, proper implementation has always been a problem and at various points in the years since independence, there have been shifts in policy focus towards single-purpose programmes such as family planning or eradication of communicable diseases, which have played their part in fewer resources being diverted to primary health care.

In 2005, the National Rural Health Mission (NRHM) was launched by the Indian Government with the aim of providing efficient primary health infrastructure to benefit the 80% of the population who live in rural areas. One of the key features in this framework was the creation of the Accredited Social Healthcare Activists (ASHAs) who are female health-care workers designed to serve as a communication link between the people in rural

areas and their health community centre. By 2012, the NRHM aimed to have approximately 400,000 ASHAs in 18 high focus states (the ones with poor health infrastructure) with the ratio of 1 ASHA per 1000 people [15]. The NRHM was also designed to create health personnel at every level with appropriate linkages amongst them. Figure 1.1 illustrates the model.

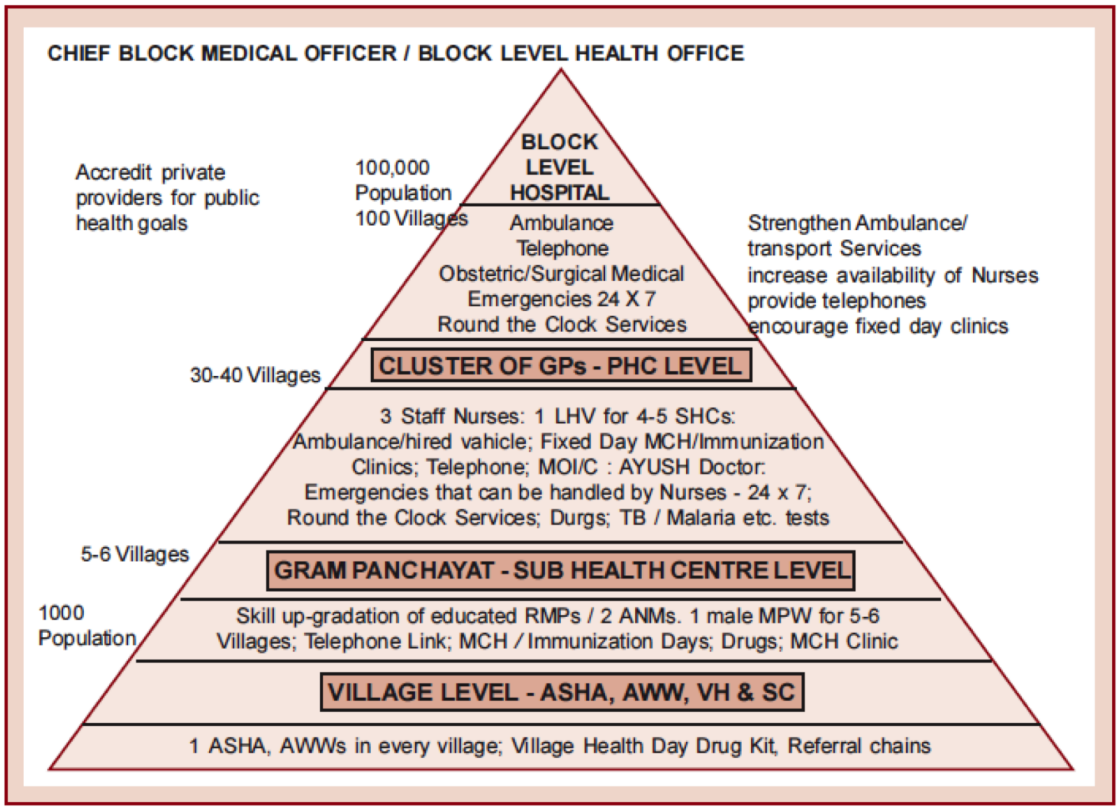


Figure 1.1 Structure of the NRHM framework adapted from National Rural Health Mission's documentation [16]. At the bottom of the pyramid are the ASHAs who form the interface between the public health system and the villages. Abbreviations: ASHA-Accredited Social Health Activist; AWW-Anganwadi worker (similar capacity as an ASHA but focuses on child development as well); SC-Sub Centre; VH-Voluntary Health workers; RMP-Rural Medical Practitioners; ANM -Auxiliary Nurse Midwife (healthcare workers at Sub-Centre level); MPW-Multipurpose workers; MCH - Maternal and Child Health; LHV-Lady Health Visitor; SHCs-Sub Health Centres; MOI/C- Medical Officer In Charge; AYUSH-Indian System of Medicine (Ayurveda, Yoga & Naturopathy, Unani, Siddha and Homoeopathy).

The public health services in rural India can be described as beginning with the *Sub-Centres* as the focal point of contact between the village community and the health system.

They are designed to cater to the healthcare needs of around 5000 people [17] and to be run by at least an ANM and a male healthcare worker [15]. Next, the *Primary Health Centres* (PHCs) operate at the block level and depending on the area and size can cater for between 30,000 and 100,000 people [17]. Generally, they are equipped with two general physicians, a staff nurse, an ANM and LHV. Recently, plans have been outlined to meet the growing demand and shortage of medical personnel in the PHCs. For instance, three staff nurses and round-the-clock services have been proposed in addition to having more doctors [15]. The next level of healthcare centre is the *Community Health Centre* (CHC) to meet the needs of between 100,000 to 300,000 people [17]. Facilities include general and specialist doctors and staff nurses. Next in the health setup are the *First Referral Units* (FRUs), which are mainly designed to provide emergency obstetric and newborn care but also have general doctors and nurses [15]. The top level public health centres at the rural framework are the District Hospitals.

In Hindi, one of India's official languages, the word ASHA translates as hope. ASHAs are often chosen by the village panchayat (a team of usually 5 respected elders of the village serving as a ruling authority) and preference is given to married, divorced or widowed women aged between 25 and 45 years. This is partly because in Indian tradition, it is common for women to move to their spouse's home after marriage. A majority of the ASHAs are educated up 10^{th} grade in school (equivalent to GCSE level). ASHAs get performance-based remuneration such as Rs.150 (£2) for every woman they bring to the PHC to get their child vaccinated or receive advice on family planning. Their role is honorary and salaries are often low, not more than Rs.1200-1500 (£18) every month. ASHAs play a major part in maternal health, child development and connecting the village's health needs to the PHC. Their familiarity with their village makes them an important aspect of the NRHM's objectives,

as well as a valuable health workforce whose capacities can be enhanced to deliver essential healthcare to villages.

1.4 Introduction to mobile health (mHealth) approaches for chronic disease management

Mobile health or mHealth is the delivery of healthcare services through mobile technology [18]. The substantial penetration of mobile phones [19] [20] even in the poorest and most inaccessible regions of the world could help to solve one of the major challenges in global health - that of the structural barriers to access [21]. The doctor-to-patient ratio is 1:20000 in rural India, as opposed to 1:2000 in urban areas [22]. Hence, there is a need to bridge the gap caused by the lack of access to quality healthcare in rural areas. Superior data quality, lower communication delays, and the option of automatic interpretation could be delivered by using electronic systems instead of traditional paper-based data collection techniques [23].

A variety of chronic disease focused mhealth systems have been implemented in the developed world for the management of long-term conditions such as type 1 or type 2 diabetes [24][25][26], medication reminders and adherence support [27][28], drug inventory management through text messages [29], data collection [30], encouragement of healthy behaviour and lifestyle change such as weight reduction [31], increasing physical activity [32] and smoking cessation [33].

Despite numerous mHealth pilot studies in recent years, limited evidence is available on large-scale adoption, efficacy, effectiveness and best strategies for engagement [21]. The basic evidence, therefore, for scale-up of mHealth programmes is insufficient with few formal outcome evaluations in LMICs [34]. There is a requirement for thorough evaluations in the

form of randomised controlled trials as opposed to mere pilot studies [21]. The successful large-scale use of mHealth may require the adoption of open health architecture for interoperability, standardised frameworks for reporting and the use of behavioural change theories [21]. mHealth applications need to be designed with the communities, health workers and patients in mind.

We described the role of ASHAs in the rural Indian healthcare in Section 1.3. If we are to scale up care for chronic diseases like CVD that require continual monitoring, this unique workforce that bridges villages and public health centre should be fully utilized. However, their lack of substantial training makes this task difficult. Hence, there is a need for enhancement of their capacity. Electronic Clinical Decision Support Systems (CDSS) offer several advantages and have the potential to create a rural healthcare workforce that can perform state-of-the art screening and management of CVD. Although contextually different, the concept of using telemedicine as a tool for capacity enhancement has been suggested as a key factor for motivating and retaining health workers in different low-resource settings (such as in Mali [35]).

1.5 Overview of thesis

Chapter 1 gives an introduction to the burden of cardiovascular disease with special emphasis on the Indian context. Brief background to the Indian primary care system, the role of ASHAs and rural PHC physicians and their capabilities for chronic disease management is also presented. The rationale for an mHealth approach and the utility of a CDSS for large-scale screening and management of CVD in India are outlined. Relevant studies from the literature are discussed.

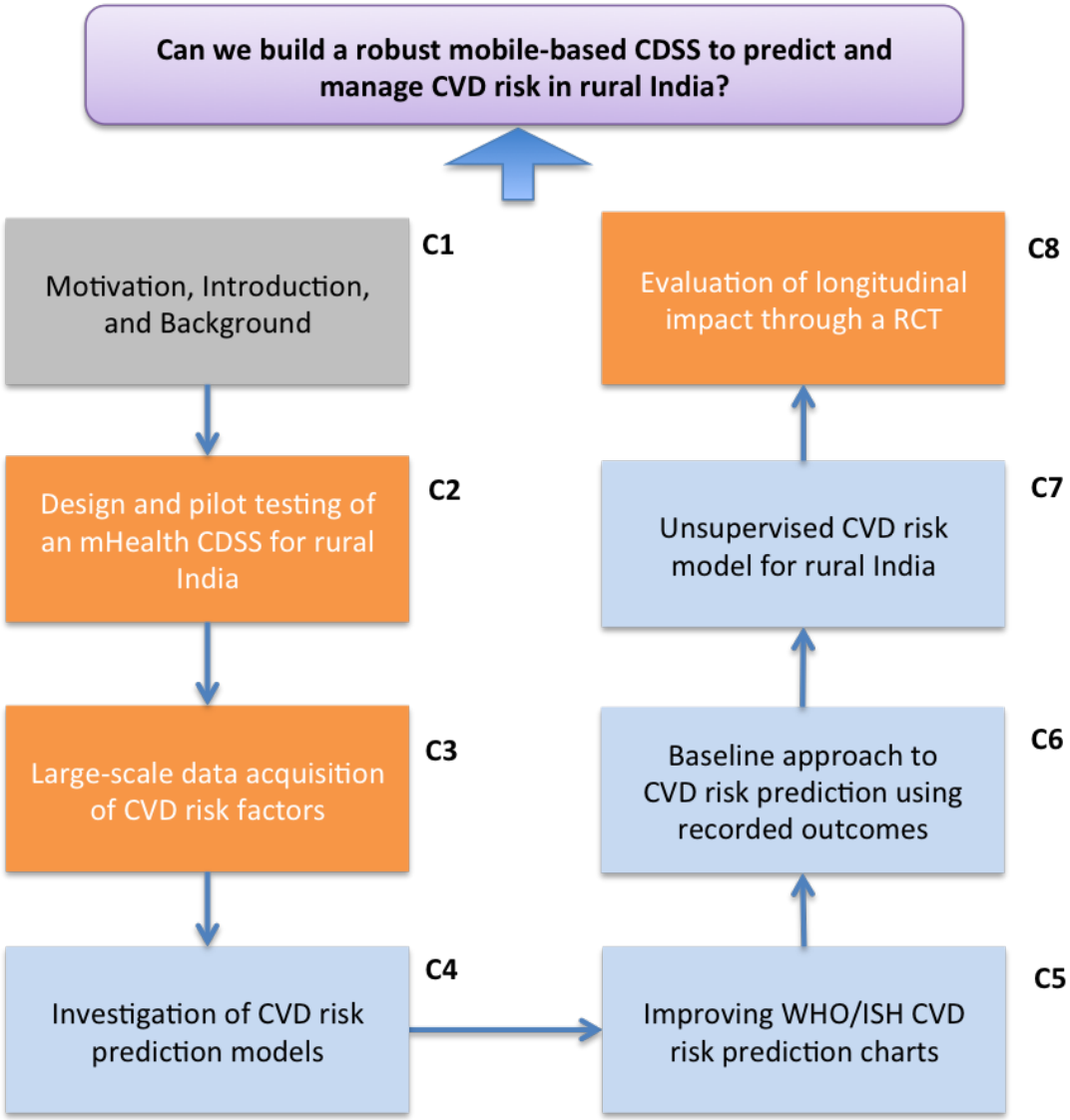


Figure 1.2 Block diagram describing the overview of this thesis. The central research question (top box, shaded in purple) is addressed through two major areas of work. The first area is centred around creation and iterative field evaluation of a mobile-based CDSS which has the elements of design, development, and data analyses. This is illustrated through the 3 boxes in orange that represents the pilot study (N_p=292), baseline study (N_b=62254), and patients from a subgroup of the first phase of a randomised controlled trial (N_{rct1}=27346) respectively. The second area is the development of data-driven models to improve algorithms for cardiovascular disease risk assessment for rural India. This is illustrated through the 4 boxes in blue. *C* stands for 'chapter'.

Chapter 2 is a discussion of the design of an mHealth tool, *SMARThealth* in a resource-limited setting for users with little or no ICT experience. Preliminary findings on the CVD risk profile, user engagement and adoption from a pilot study of 11 ASHAs, 3 PHC physicians and 292 participants across villages in Andhra Pradesh are presented.

Chapter 3 is a description of *SMARThealth* with minimal modifications in order to perform large-scale data collection. The process behind data acquisition from all eligible participants across households of 54 villages, constituting 62194 participants, is discussed. Different CVD risk factors including the CVD risk distribution are presented. Multivariate data is visualised, which revealed underlying patterns in the data structure as an exploratory step in data analysis.

Chapter 4 is a description of risk prediction models for CVD, which form the core of the mobile-based CDSS. A technical overview of the major existing risk prediction models in the literature is presented. Datasets used in this thesis and an outline of improvements to CVD risk prediction models proposed in the next few chapters are then discussed. A comparison of the World Health Organisation/International Society for Hypertension (WHO/ISH) risk prediction charts for WHO South East Asian Regions D (SEAR-D) with the existing benchmark algorithm, the Framingham Risk Score is presented based on data from rural India.

Chapter 5 is the presentation of detailed analysis of the sole algorithm prescribed for CVD risk assessment in India, namely the WHO/ISH CVD risk prediction charts for SEAR-D. These charts were programmed into the mobile-based CDSS. The chapter begins with a presentation of essential technical background on machine learning techniques. Subsequently, the focus is shifted to two important contributions on risk prediction using the WHO/ISH charts in India: (1) Evaluation of the clinical implications of using the WHO/ISH charts

on a rural Indian population (analysis performed on a 1066 patient database with recorded cholesterol measurements from Andhra Pradesh); and (2) Development of a patient-specific point-of-care algorithm to determine the benefit of a total cholesterol test during risk assessment using WHO/ISH charts.

Chapter 6 is an investigation of a machine learning approach towards CVD risk prediction. Feature selection is performed to investigate highly predictive features on benchmark data from the Framingham study, which has recorded outcomes. An approach to recalibrating the obtained prediction model for rural India is discussed.

Chapter 7 is a discussion of an alternative approach to identification of high-risk patients through unsupervised learning. Clustering is performed with two large databases from India (N_b=62194; without outcomes) and the USA (N_f=3040; with outcomes), and the performance is evaluated.

Chapter 8 is a presentation of a randomised controlled trial design for evaluation of the effectiveness and clinical impact of the CDSS tool, *SMARThealth*. This is amongst the largest mHealth studies conducted in lower and middle income countries to-date. The production-ready CDSS for the RCT is presented, along with customisation of an open-source medical record system for use in rural India. Intermediate results from the first phase of the RCT are analysed and preliminary findings on the change in blood pressure levels on those patients receiving the intervention, is presented.

Chapter 9 is a summary of the work presented in this thesis. An overview of future work is also discussed.

Chapter 2

Building a mobile-based clinical decision support system

2.1 Introduction

A recent review on mHealth for LMICs has concluded that despite the potential of a variety of applications in non-communicable disease care, mHealth is currently dominated by behaviour change interventions (e.g. smoking cessation) [36]. In this chapter, the design and pilot testing of a multifaceted healthcare worker intervention utilising an mHealth platform are presented. The platform comprises a clinical decision support system called *SMARThealth*, for CVD risk assessment and management in rural India.

For this pilot study, data on CVD and associated risk factors were collected from villages surrounding 3 primary health centres along the West Godhavari district in rural Andhra Pradesh. This was achieved through an mHealth system comprising a client-side mobile application and server-side electronic medical record system. The mobile application was compartmentalised into four simple steps that enabled health workers to conveniently assess rural participants for CVD risk and disseminate appropriate recommendations. Table 2.1

describes the four-step process of data acquisition through the mobile application. Section 2.2 presents the process of building a CDSS. Implementation of the CDSS through the creation of a mobile application suitable for rural India, is discussed in Section 2.3. Field evaluation of this system was performed as detailed in Section 2.4. The outcomes recorded focused on a preliminary evaluation of the *SMARThealth* tool for utility, effectiveness and acceptability by the ASHAs and community participants in this setting in order to inform large-scale evaluation. Results are presented in two parts: (1) Section 2.5 details the analyses of data collected and discusses risk factors and CVD risk profile as determined from the studied population; (2) Section 2.6 provides a quantitative evaluation of the mHealth platform based on analysis of user interactions with the application. The lessons from the pilot study are then discussed in the context of scaling up the platform for a large RCT in order to investigate clinical impact.

The work presented in this chapter aims to mitigate two major problems with current mHealth solutions. Firstly, technology-driven global health solutions have conventionally been imposed on low-resource settings, with little knowledge of the latter. To overcome this, an iterative design process considering the background and involvement of the local health workforce is presented, and usage patterns from the end-users are analysed to provide detailed insight on workflow integration. Secondly, targeted mHealth solutions are ad-hoc and usually lack interoperability. By utilising open-source solutions in the mHealth system architecture, we aim to mitigate the second problem.

2.2 Components of the clinical decision support system

2.2.1 Computing 10-year absolute CVD risk

The WHO/ISH provide colour-coded charts [37] that predict the 10-year risk of fatal or non-fatal cardiovascular event (myocardial infarction or stroke) in different epidemiological

Table 2.1 Table showing the 4-step process designed in the CDSS for collecting patient data and providing decision support. Terminology used: PVD - Peripheral Vascular Disease, SBP - Systolic Blood Pressure, DBP - Diastolic Blood Pressure, TC - Total Cholesterol, LDL - Low Density Lipoprotein, HDL - High Density Lipoprotein, TG - Triglycerides, Patient ID - Patient Identifier

Step 1 (Demographics)	**Step 2** (Medical history)	**Step 3** (Risk factor acquisition)	**Step 4** (Decision support)
Age Gender Location Patient ID Name	Past history of Myocardial infarction/Angina Past history of Stroke Past history of PVD Past history of Diabetes Family history of Myocardial infarction/Angina Family history of Stroke Family history of Diabetes	Blood Pressure Blood Glucose (fasting/random) Cholesterol (TC,HDL,LDL,TG) Height Weight	Recommendations (Smoking cessation, Nutrition and Lifestyle modifications), Next Visit (CVD risk screening, Diabetes screening physician referral), Medication (Blood pressure lowering, Lipid lowering, Anti-platelet therapy), Tagets (SBP, DBP)

sub-regions of the world. The WHO/ISH charts for SEAR-D, which includes India, were employed for risk prediction in our mobile CDSS. Depending on the availability of cholesterol information, Low Information (LI) or High Information (HI) versions of the CVD risk charts (illustrated in Section 4.2.5 of Chapter 4) can be used with the latter including total cholesterol as a predictor. Detailed descriptions of the WHO/ISH charts are available in Chapter 5. The colour-coded ranges in the WHO risk charts indicates five levels of CVD risk for different values of risk factors [37]. Information necessary for risk estimation includes:

- Presence or absence of diabetes
- Gender
- Age
- Smoking status
- Systolic blood pressure (SBP)
- Total blood cholesterol (TC), if known.

2.2.2 Estimating sub-conditions for management

Subsequent to the computation of a 10-year CVD risk for a given set of risk factors, the *SMARThealth* mobile application offered support for management of the participant's risk and further follow-ups. In order to arrive at appropriate recommendations, the tool calculated sub-conditions (as listed in Table 2.2) that indicated high-risk conditions or the presence of elevated glucose levels (impaired fasting glucose or IFG), high BMI (indicating obesity), or hypertension.

Table 2.2 Computation of essential sub-conditions based on patient's assessment. This is performed as a precursor to recommendations for CVD management; Terminology - BMI stands for Body Mass Index and is the weight (in kg) divided by the height squared (in metres)

Clinically High Risk	**Impaired Fasting Glucose**	**Weight**	**Hypertension**
SBP≥160 mmHg or DBP≥100 mmHg or TC≥320 mg/dL or LDL≥240 mg/dL or TC/HDL>8	Fasting blood sugar level between 110 mg/dL and 126 mg/dL or history of diabetes	Obesity if BMI≥30 Overweight if 25≤BMI<30	SBP≥140 or DBP≥90 or history of hypertension

2.2.3 Management and treatment of CVD

After computation of the CVD risk score and sub-conditions for assessing high risk, the *SMARThealth* CDSS was designed to offer point-of-care (POC) decision support to both ASHAs and physicians for the management of CVD. This included three categories - general recommendations, referral, and treatment recommendations, based on international and national guidelines [38] [39]. General recommendations are intended for controlling or modifying behavioural risk factors including smoking, alcohol consumption, physical inactivity, and diet/nutrition. Except for the detailed advice on smoking cessation (which is given only to smokers), other recommendations are displayed to all users. A referral indicates the need for consulting a primary care physician, which is also the outcome for 'next visit' flags for absolute risk and diabetes screening. Treatment recommendations were available only to physicians since ASHAs are not authorised to prescribe medication. The CDSS's output for different CVD risk profiles are summarised in Table 2.3.

2.2.4 Validating management guidelines for CVD

The prediction and management applications were stringently validated in two stages, as described below.

Table 2.3 Referral and medication recommendations for the three categories of CVD risk considered in this work. Terminology - IFG stands for impaired fasting glucose

Risk category		**Referral**		**Medication**
		CVD risk screening	physician referral	
High Risk	≥30% risk or past CVD history	every 3-6 months	Yes	BP lowering therapy Lipid lowering Antiplatelets
Intermediate Risk	20 to <30% risk	every year	Yes	BP lowering therapy (if BP≥ 140/90) Lipid lowering (if Diabetes is present or SBP≥160 mmHg or TC>200 mg/dL or LDL>120 mg/dL)
	10 to <20% risk	every 2 years	Yes if (Diabetes or IFG is present) or (history of diabetes and SBP≥160 mmHg)	BP lowering therapy (if Diabetes is present and BP≥140/90 or SBP≥160 mmHg) Lipid lowering (if Diabetes is present or SBP≥160 mmHg)
Low Risk	0 to <10% risk	every 5 years	Yes if (Diabetes or IFG is present) or (history of diabetes and SBP≥160 mmHg)	BP lowering therapy (if Diabetes is present and BP≥140/90 or SBP≥160 mmHg) Lipid lowering (if Diabetes is present or SBP≥160 mmHg)

Code validation

Each of the input and output variables programmed in the CDSS was tested initially using a large de-identified dataset of 200 patients from the Andhra Pradesh Rural Health Initiative (APHRI). The CVD risk prediction and management algorithm detailed in the previous sections, was implemented by two physicians using the SPSS statistical package. The author of this thesis, a biomedical engineer, coded the algorithm as an Android application in Java. Both implementations were tested on a 1000-patient primary health care dataset with cardiovascular disease risk factors. Over multiple iterations, coding errors were spotted whenever outputs disagreed between the two implementations and adjustments were made to the algorithm for robustness. Outputs were compared between the physician coded

implementation and the non-physician coded implementation until a near-perfect correlation with clinically insignificant errors was obtained.

Independent physician validation

De-identified data for 100 randomly selected individuals enrolled in the APHRI study was assessed by a physician researcher not involved in the CDSS development. She assessed the 10-year CVD risk of each individual using the WHO/ISH risk charts and manually reviewed the CVD management advice (BP-lowering, anti-platelets and lipid-lowering medications) and targets for patients already on BP-lowering and Lipid-lowering drugs. Agreement was assessed between the CDSS programmed as an Android application, and the independent physician's recommendations, and adjustments were made subsequently to the *SMARThealth* tool's management advice and referral sections based on the physician's recommendation.

2.3 Android application development

2.3.1 Designing a mobile health application for rural India

An important factor in the success of any technology-based intervention depends on how well it can fit within the needs of the users and the environment [40]. In our context of managing CVD in resource-constrained settings, the challenge and opportunity was to design an appropriate tool that could be easily used by minimally trained health workers (who were the end-users) as well as fit the needs of local primary care physicians and community participants.

Conventional methods of design and development rely on traditional "requirements engineering" [41], where the emphasis is on preparing thorough documentation with clear specifications prior to the development of a proposed solution. This is often the first phase in the "waterfall" software development model which relies on a pre-planned sequential design

procedure [42]. Another popular practice in software engineering is the agile development methodology which emphasises iterative development with constant feedback from the stakeholders, whereby requirements emerge along the process [43]. This has the advantage of being able to evolve according to changes in a dynamic environment.

The requirements of our end-users compounded with the uncertainties in technical infrastructure (such as the availability of uninterrupted 2G/3G internet connectivity) could not be gauged thoroughly through a needs assessment. Therefore, the agile development approach was followed in our study to design prototypes iteratively and elicit feedback from the end-users at each step. However, our approach differed from conventional practice since the end-users had little or no experience of using information and communication technology and could not drive the design. To overcome this, firstly we employed a multidisciplinary team comprising an engineer, a local physician, a sociologist, and an expert physician to balance the *system requirements* (for designing features for an effective intervention by understanding the local clinical practice for conveying recommendations for managing CVD risk, or assessing the accuracy of ASHAs asking questions related to medical history) with *user requirements* (for instance, observing if the ASHAs can compute the year of birth based on age). Secondly, we performed phases of prototyping and user assessment (as illustrated in Figure 2.1) whereby the multidisciplinary team evaluated the end-user interactions with the mobile application through observations and post-procedure interviews.

2.3.2 Key features

Some of the key features that were incorporated through our agile approach included:

- **One touch navigation**

  Since the end-user base had little or no experience, touch errors were common and frequent. The mobile application's design was such that ASHAs could navigate

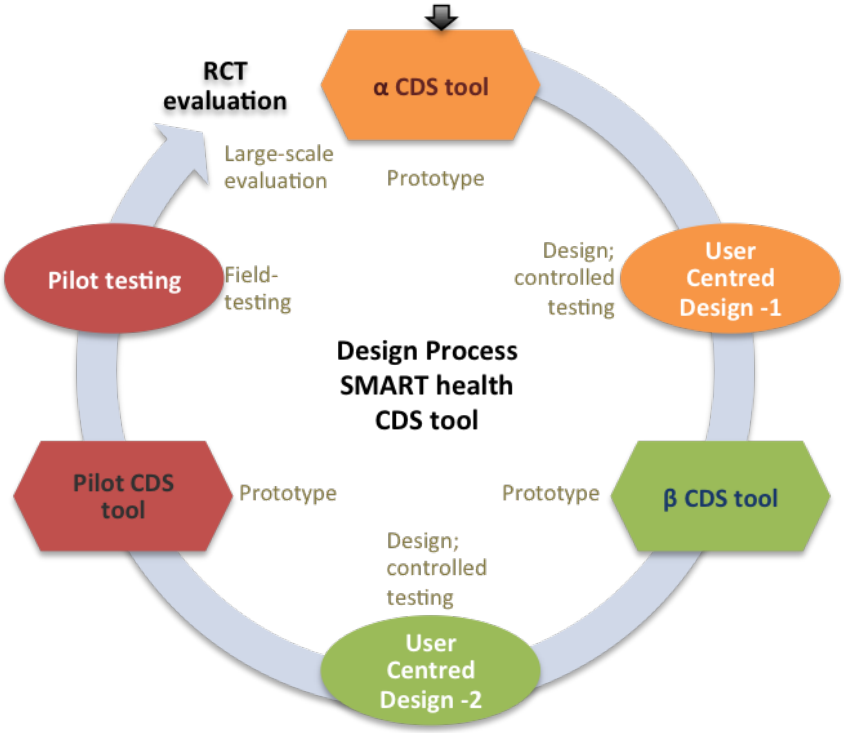


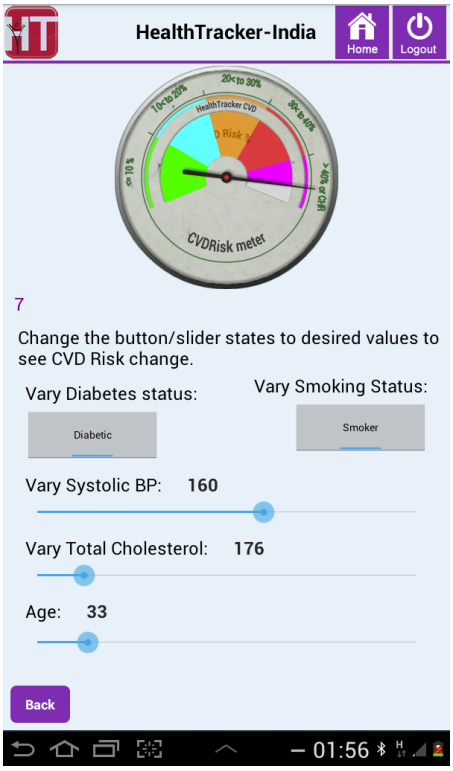
Figure 2.1 Phases of design and development following a user centred design approach, where the aim was to bring out issues centred around the end-user's interaction with a prototype without placing any explicit demands on them. The process diagram shows prototypes in the form of an alpha (α) and beta (β) CDSS, each followed by a user centred design phase. This led to the development of a tool for the pilot study, which upon further improvement would be suitable for a randomised controlled trial.

through the majority of the content using a single button. Care was taken to ensure idiosyncrasies of the mobile-based device that interfered with workflow were minimal. For example, in certain cases the virtual keyboard was minimised subsequent to user input.

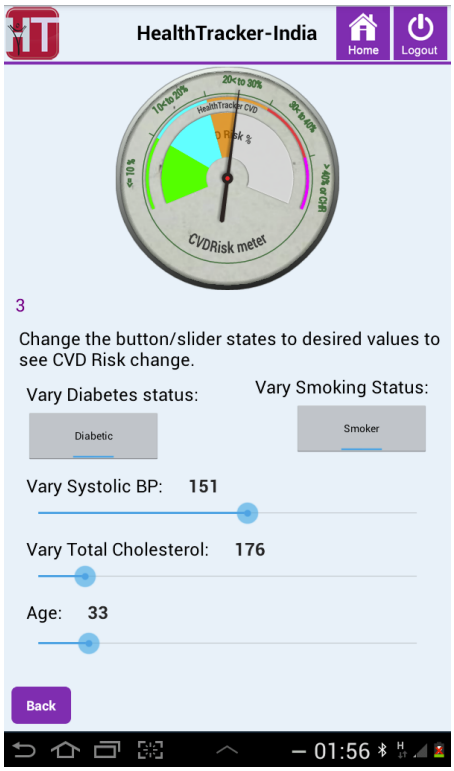
- **CVD risk projection meter**

  The application was developed with an embedded visual risk projection meter (as shown in Figures 2.2a and 2.2b) to convey the meaning of CVD risk to the patient in an understandable way. For example, Figure 2.2a illustrates the case of a 33-year old smoker with diabetes. The participant has a blood pressure value of 160/89 mmHg, total cholesterol of 176 mg/dL and his risk of developing CVD over a 10-year period is

shown to be high. If the participant reduced his blood pressure alone by approximately 10 mmHg (as shown in Figure 2.2b), his 10-year risk of developing CVD is reduced substantially from above 40% to between 20% and 30%. Thus the cause and effect of each risk factor for a particular patient can be graphically explored in an interactive manner. This is intended to help patients visualise the effect of controlling risk factors and was designed to encourage adherence to medication and a change in behaviour affecting modifiable risk factors, such as smoking.



(a) Present risk factor levels for the patient



(b) The projected 10-year risk if the patient reduces his blood pressure by approximately 10 mmHg

Figure 2.2 CVD Risk projection meter

- **Event calendar/age entry**

  It was found that some of the participants did not know their date or even year of birth accurately. This was due to factors such as date of birth being mandatory for

registration only after the Indian Registration of Births and Deaths Act, 1969 [44]. Therefore, the mobile application was programmed with methods to increase the accuracy of the estimated age; a vital parameter in the risk assessment algorithm. A comprehensive list of well-known historical events (for example, Indian Independence year, 1947) could be retrieved in order to help the patient decide how old they were at that point in time. Also, for female patients, specific questions (mostly related to maternal health) that had a high probability of narrowing down the exact age were used. The list of 'events' is described in Appendix B.

- **Information buttons**

  These buttons were embedded in the standard risk assessment procedure and intended to disseminate essential information for a particular question the user wanted to know more about. For instance, the button could be used to recap the protocol for blood pressure measurement or retrieve a list of commonly used drugs to see if the patient had a history of taking any of them.

- **Review patient**

  The review patient feature offered the option to review existing patients who had been screened previously by that user. This was especially motivated by the recommendations of a physician who intended to use a staff nurse to screen patients before they had an appointment with him. Using the Review Patient feature, the physician could assess and review the participant's CVD risk in an efficient way, thereby promoting greater ease of adoption of the tool in the primary care setting.

- **Accuracy of data entry**

  To ensure correctness of data entry and avoid transcription errors, automated on-screen checks were performed at the end of each step that alerted the user if the entered data was not clinically valid. Furthermore, a validated Blood Pressure (BP) device

(Stabil-O-Graph [45]) was programmed to interface with the tablet via Bluetooth and transmitted three BP measurements to the tablet. Blood glucose was measured using the OneTouch Ultra2 glucometer which did not have the desirable Bluetooth functionality. This was because the recurring cost of test strips was prohibitively high for a Bluetooth-enabled glucometer. All data entered in the CDSS could be corrected or inputted manually prior to proceeding to the next step of the risk assessment procedure.

2.4 Field testing

The mobile based CDSS was field tested by 11 ASHAs who performed door-to-door screening of community participants in 3 villages in rural Andhra Pradesh, India. Each ASHA was designated to screen approximately 20 participants in one month for high risk and a convenience sampling method was followed, whereby participants could be selected based on their accessibility and proximity to the ASHAs. The ASHAs were trained for one week to use the CDSS and tablet. The data from each participant screened by the ASHA were uploaded to a secure server hosting an electronic medical record system (OpenMRS [46] [47]). Decision support on treatment was available to physicians, who used the CDSS as a standalone system to screen and manage patients visiting their clinic. The entire mHealth infrastructure was designed to be interoperable and leveraged the Sana Mobile platform, an open-source telemedicine framework [48] [49].

The CDSS was built as a mobile application for devices running Android 4.0 and above, and supported both Telugu (local language) and English. The application was optimised to run on a higher-end version of a low-cost tablet (£100) for use by ASHAs, and a Samsung Galaxy 7-inch tablet (£160) to be used by physicians. Low-cost tablets were preferred for two reasons. The first was to assess if they could perform adequately in our setting and

inform the cost-effectiveness of the intervention. The second reason was the high healthcare-worker:doctor ratio in resource-constrained regions which required cost-effective tablets (since they can be replaced more easily if damaged or stolen).

A four point Likert scale [50] was used to evaluate usability at the end of every risk assessment performed by the ASHAs. To understand user behaviour and pattern of interaction with the mobile application, an analytics framework was built within the tool. It recorded the data and timestamp for every click made by the ASHAs as they performed CVD risk assessments. For ease of comparability of usage patterns and to highlight diversity in trends amongst different end-users, three distinct ASHAs were chosen based on the total number of assessments they performed and the years of experience they had in performing their role. Each of the chosen three ASHAs was identified through interviews as being primarily a Telugu speaker, and claimed to be able to read and write English at secondary school level.

The variability in end-user usage was analysed in this study to gauge the performance of ASHAs. This was performed in order to identify the stage at which additional support or training was needed, if at all necessary. Data from the three ASHAs, representing all the ASHA's performance, were analysed individually to obtain an estimate of their mean procedure time over the course of the pilot study. Bootstrapping was performed considering all samples available up until and including that procedure. This was performed to mitigate bias towards uneven sampling. For example, if we are to estimate the mean completion time and confidence interval (CI) for an ASHA who has performed 10 risk assessments, her procedure times until and on the 10^{th} procedure would be taken as our sample for bootstrapping. The 95% CI for the mean was subsequently estimated and plotted individually for the three ASHAs. This is useful to understand how variable the ASHA's own procedure times can be and with sufficient samples (of risk assessment procedures), we can reliably estimate how long an ASHA can take and how much training is needed until she is proficient at using the

CDSS (which can be shown by a narrow confidence interval).

The ASHAs manually entered the glucose values into the tablet, a process susceptible to errors. However, to assess the extent of errors, the blood glucose values stored in the glucometers were downloaded and compared with the values entered by the ASHA in the tablet during risk assessment. Care was taken to ensure that the glucose readings being compared were taken at the same time. This was performed by comparing the timestamps from the analytics framework in the CDSS and those from the glucometer's memory.

In-depth qualitative interviews with the ASHAs and focus group discussions were also undertaken to identify barriers to the adoption of mobile technology. These are detailed in an article written by this author [51] (see Appendix C).

Ethics The study was approved by the ethics committee of the Centre for Chronic Disease and Control, India and the University of Sydney, Australia. Informed, written consent was obtained from all participants contributing data in the study.

2.5 Results - Risk factors and CVD risk profile

A total of 227 participants (mean age of 51.4 $\pm$13.1 years) were screened by 11 ASHAs, while 3 PHC physicians independently used the mobile clinical decision support tool to screen 65 in-patients (mean age of 55.3 $\pm$ 11.7 years) who visited their clinic (total N=292).

Tables 2.4 and 2.5 summarise the data collected in the field. Statistical significance testing was performed using the two-sample, two-sided t-test for continuous variables and

the Wilcoxon signed-rank test for discrete variables.

Table 2.4 Baseline characteristics of all screened participants (N=292). Statistically significant differences due to gender are italicised. All characteristics except blood pressure were significantly different. Abbreviations- BP indicates blood pressure; BMI stands for Body-Mass Index

Feature	**Male**	**Female**	P value
N	39.7% (116)	60.3% (176)	
Age, mean±sd, (years)	55.4±13.9	50.2±11.8	<0.001
Current Smoker,%(n)	34.5% (40)	5.1% (9)	<0.001
Systolic BP, mean±sd, (mmHg)	129.4±18.5	129.1±22.8	0.90
Diastolic BP, mean±sd, (mmHg)	82.1±11.4	79.3±12.2	0.052
BMI, mean±sd, (kg/m^2)	23.4±4.0	24.7±5.0	0.026
Elevated Blood glucose, %(n)	37.1% (43)	25.6% (45)	0.036

Table 2.5 Medical history of participants screened using the CDSS. Differences on account of gender that were statistically significant (p<0.05) are italicised. Smoker indicates the person is either a current smoker or quit smoking within the last 12 months.

Parameter	**Male** (N_m=116) n (%)	**Female** (N_f=176) n (%)
Past history- Heart attack/angina	13 (11)	30 (17)
Past history- Diabetes	26 (22)	18 (10)
Past history- Stroke	3 (2)	1 (1)
Past history- Peripheral vascular disease	6 (5)	17 (10)
Family history- CVD	8 (7)	16 (9)
Family history- Diabetes	19 (16)	19 (11)
Medication history- BP	32 (27)	41 (23)
Medication history- Lipid lowering	3 (2)	0 (0)
Medication history- Anti-platelets	5 (4)	3 (2)

Figure 2.3 illustrates the pairwise correlation for risk factors mentioned in Table 2.4. Only systolic and diastolic blood pressure showed substantial correlation (0.74).

The risk profiles of all participants were analysed using the low information WHO/ISH LI charts. Figure 2.4 shows the distribution of a 10-year risk of developing CVD for the

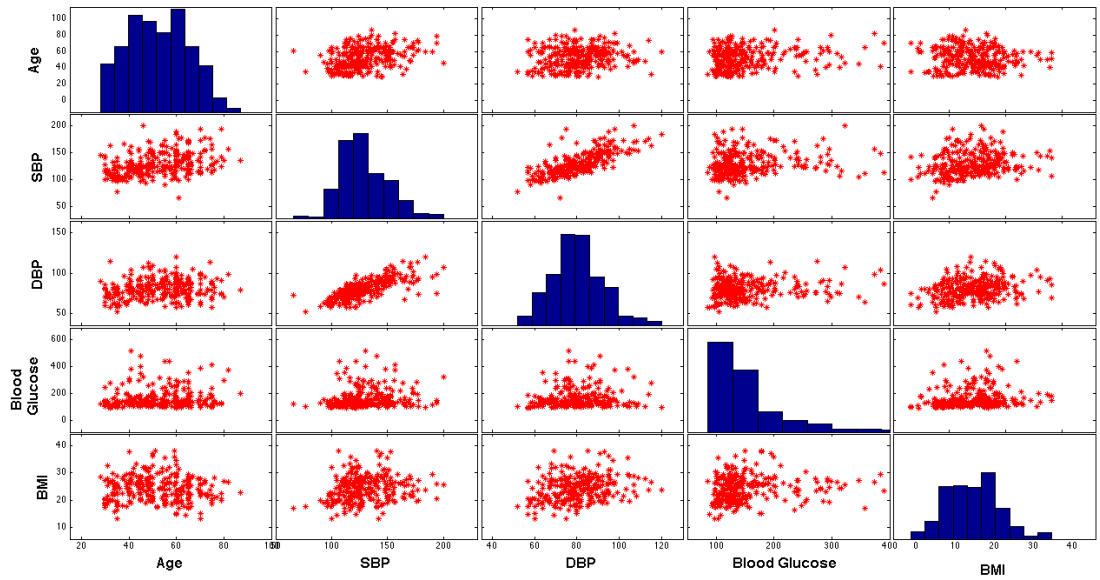


Figure 2.3 A matrix of scatter plots showing pairwise correlation amongst the risk factors obtained from 292 participants in the pilot study. The leading diagonal illustrates a histogram of the risk factor. The systolic and diastolic blood pressures show high correlation (0.74) as expected. Age and BMI have no relationship (marginally negative value of -0.08) for the pilot participants.

participants for the study participants. A majority of participants were either at low risk (<10%) or high risk (due to a previous incidence of CVD, or had a clinically high condition as previously defined in Table 2.2, or both). Physicians who used the CDSS to screen in-patients prescribed recommendations and medication for lowering CVD risk. However, for those who were already on BP medication, it is worthwhile noting that female patients did not meet their SBP and DBP targets in comparison to male patients (as illustrated in Figure 2.5). This could be due to a variety of reasons including greater mobility of male patients and hence ability to receive continual medication from the PHC, physiological differences and efficacy of the given class of BP medication, or compliance. This demonstrates the necessity for a large-scale intervention with longitudinal and frequent BP measurements, in addition to medication compliance data and other system-related confounders, if the assessment of those reaching BP target levels is to be robust.

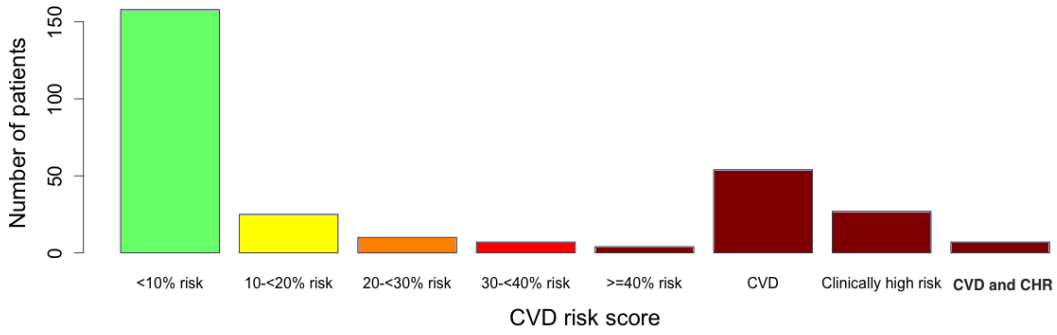


Figure 2.4 Distribution of CVD risk scores in the pilot population (N=292). CHR denotes clinically high-risk. 34% of screened participants had a CVD risk score greater than 30% or had CHR or CVD or both.

2.6 Results - Mobile analytics for gauging user behaviour and interactions

Although physicians used the mobile application, results presented in this thesis focus mainly on usage by the ASHAs. The 11 ASHAs in this study had a mean age of 31.5 $\pm$ 5.1 years and mean time of 4.6 $\pm$ 2.2 years as an ASHA. The three ASHAs chosen for detailed usage analysis, had a mean age of 29 ± 2 years and had the experience of using a mobile phone prior to the commencement of this study. The three ASHAs will be referred to as *npB1*, *npM1*, and *npL1* and they had 3, 7, and 4 years of experience in working as an ASHA and performed 28 (highest number in this study), 20 (median number of assessments in this study), and 16 (lowest number of assessments amongst ASHAs in our study) risk assessments respectively.

2.6.1 System efficiency

The overall time taken for screening the community participant's CVD risk over the duration of the study is shown in Figure 2.6. The median time for all ASHAs was 21:10 (read in terms

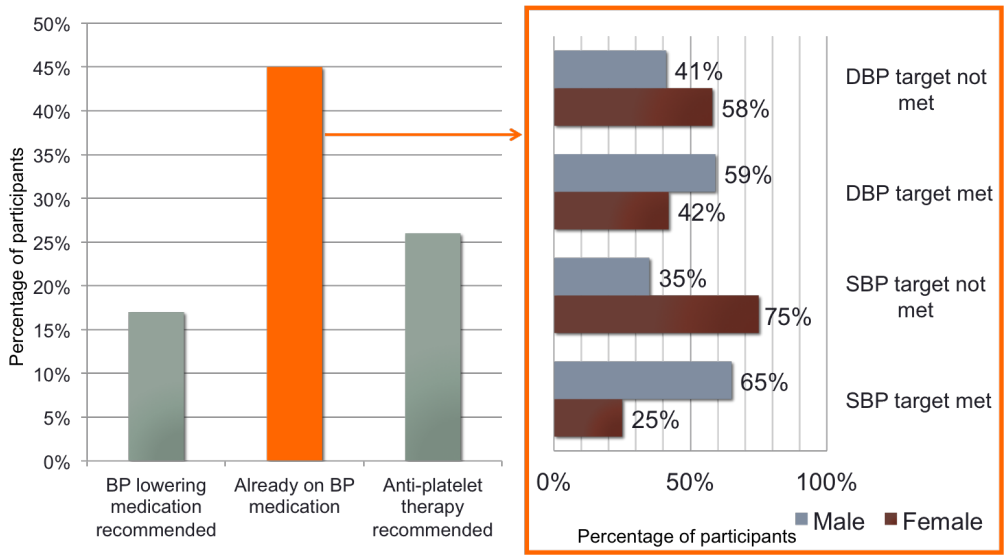


Figure 2.5 Medication and BP targets as calculated by the CDSS for in-patients screened by the physician (N_{phy}=65). SBP and DBP targets were met by a greater proportion of male patients than female patients.

of minutes:seconds) with an Inter-quartile Range (IQR) of 14:08 (minutes:seconds). ASHAs *npB1*, *npM1*, and *npL1* took 27:28 (IQR 14:05), 24:20 (IQR 12:12), and 33:53 (IQR 32:41) respectively. A decreasing trend in completion time as users performed more procedures is observed.

The distribution of time taken for each step of the CVD risk assessment for all ASHAs is shown in Figure 2.7. Step 3 (risk factor acquisition) took the longest during data collection and its extensive spread observed from the distribution shows a large variation (IQR 11:22) between ASHAs in completing that step.

The number of times the ASHA chose Bluetooth transmission over manual transmission (or the Bluetooth BP device usage rate) was analysed and is shown in Figure 2.8. Only one transmission from the BP device was needed for three BP measurements in each CVD risk assessment procedure. From Figure 2.8, ASHA *npB1* had a usage rate of 61% (17/28 procedures) with mean BP acquisition time of 05:20 (standard deviation or σ = 03:36) while

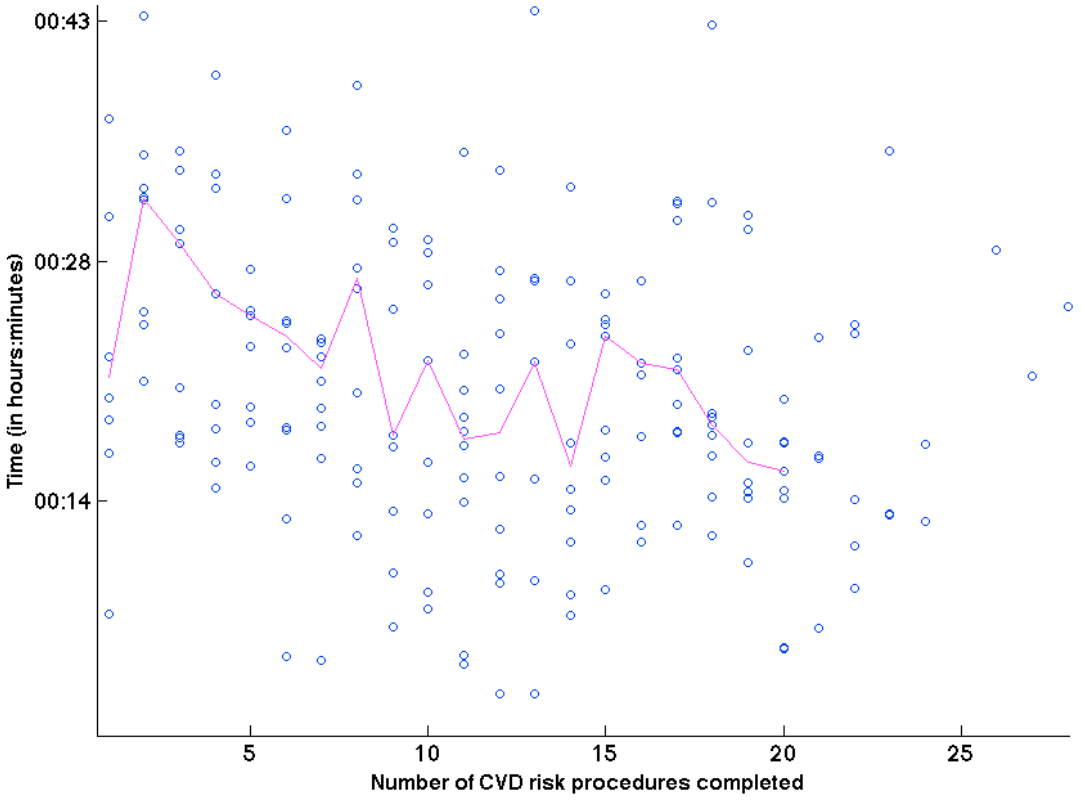


Figure 2.6 Graph illustrating the total CVD risk assessment procedure time over the number of procedures performed by the ASHAs. The median time for completion of a risk assessment was 21:10 (read as minutes:seconds). The median time for all ASHAs was taken at every procedure and the resulting trend suggests a decrease in total procedure time as more CVD procedures were performed. The first procedure has a lower median time and this may be because it was performed immediately after the training phase for the ASHAs.

npM1 had a usage rate of 50% (10/28) with mean BP acquisition time of 07:06 (σ = 02:36). ASHA *npL1* had a usage rate of 37% (6/16) with mean BP acquisition time of 08:24 (σ = 02:56). The ASHA who used Bluetooth the most (*npB1*) for transferring BP readings had the lowest mean BP acquisition time. However, the ASHA with most experience (*npM1*) showed more consistency when acquiring BP.

Before the 10th procedure, Bluetooth usage rate was 40% for *npB1*, 50% for *npM1*, and 10% for *npL1* while after the 10th procedure (midway between their 'assigned' target of 20

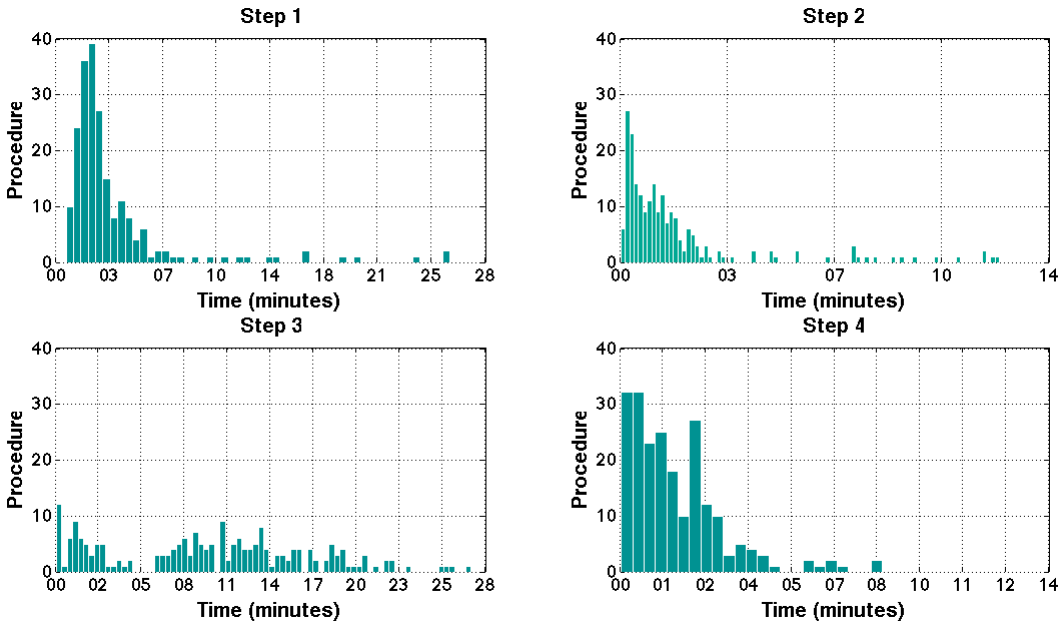


Figure 2.7 Assessment of individual step times, broken down as Step 1- Demographics with median time of 02:37 (IQR 02:16), Step 2- Medical History with median time of 01:08 (IQR 01:35), Step 3 - Risk factor acquisition with median time of 10:33 (IQR 11:22), and Step 4 - Decision support 01:30 (IQR 01:53).

procedures as previously described in Section 2.4), the usage rates were 65%, 44%, and 80% respectively.

2.6.2 User variability

From Figure 2.9, we observe that *npM1* is consistent in her mean time to perform an assessment and has the most narrow CI [21:01, 28:27]. *npL1* has the widest CI [31:58, 57:58] for the estimated mean procedure time towards the end of the study. The three ASHAs show less variability in procedure time over the course of the pilot study (when they complete more risk assessments).

In the initial stages of the pilot study (for instance, at the end of 5 procedures completed by each ASHA), the bootstrapped estimate of the mean procedure times for all ASHAs are high with wide CIs. ASHA *npL1* has the largest estimate with mean time 53:02, 95% CI

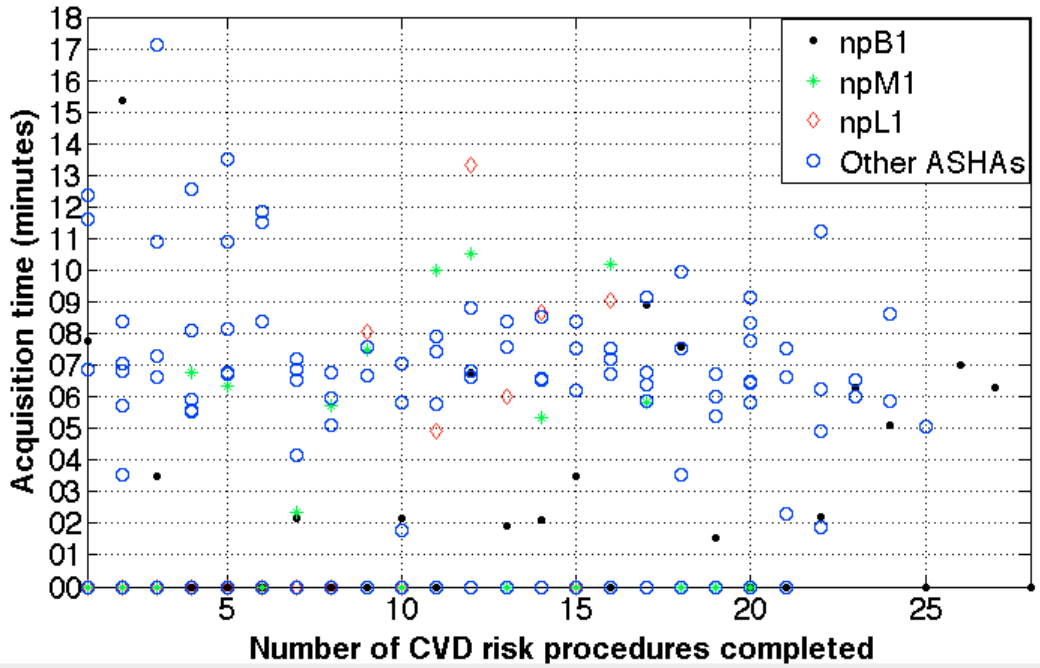


Figure 2.8 Plot of BP acquisition times over procedures performed on the course of our pilot study. Acquisition times at 0 indicate Bluetooth transmission was not attempted. Overall Bluetooth BP device usage rate was 55% and overall median acquisition time was 06:50 (IQR of 02:33). Four out of 219 procedures were not considered as the risk assessments were performed non-sequentially by the ASHAs.

[32:36, 01:24:35]. ASHAs *npM1* and *npB1* have comparable estimate of mean times but the latter has a much wider 95% CI (*npM1* - mean time 30:33, 95% CI [26:30, 32:45]; ASHA *npB1* - mean time 35:45, 95% CI [20:00, 63:28]).

At the end of 10 procedures (or approximately mid-way through the assigned target of 20 assessments for the ASHAs), the estimated mean time and CI continues to decrease (*npL1* – mean time 43:31, 95% CI [31:05, 01:04:19]; *npM1* – mean time 26:28, 95% CI [21:50,30:29]; *npB1* – mean time 35:18, 95% CI [24:34, 55:45]).

When 15 procedures have been completed by the ASHAs, the estimated mean stabilizes while CI continues to become narrower (*npL1* – mean time 41:05, 95% CI [30:56, 56:24];

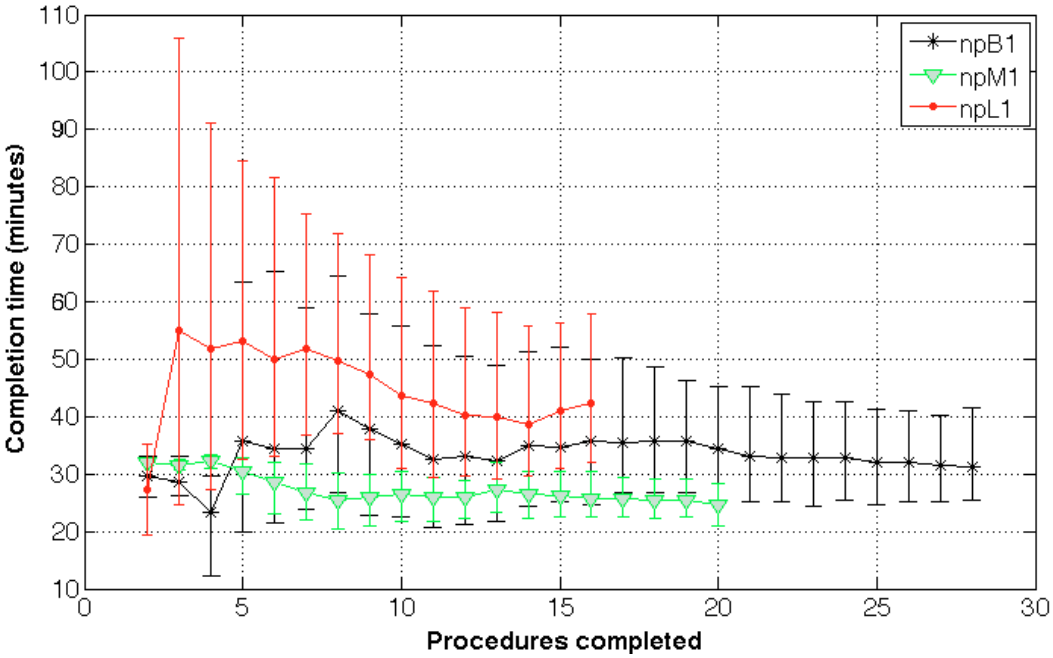


Figure 2.9 Estimate of the mean procedure times with 95% confidence intervals for ASHAs *npB1*, *npM1*, and *npL1* over successive procedures performed (for the duration of the pilot study). ASHA *npM1* has the narrowest CI and least estimated mean procedure time while *npL1* has the widest CI and highest estimated mean procedure time at the end of the study. The estimate of 95% CI and mean procedure time was obtained using a bootstrapped sample that included risk assessments performed up until and including that time period in the study. This was performed to account for uneven sampling bias (for example, at the 5th procedure, procedures 1 to 5 were taken as samples and bootstrapped). After ASHA *npL1*'s 13th procedure, her estimated mean procedure time increases which suggests that she took much longer to complete the last two risk assessment procedures.

npM1 – mean time 26:18, 95% CI [22:34,30:22]; *npB1* – mean time 34:43, 95% CI [25:07, 52:08]).

At the end of the pilot study, *npL1* has an estimated mean time 42:15 with 95% CI [31:58, 57:58] that is compared to the last milestone (15 procedures) and there has not been an appreciable increase in the number of procedures the ASHA had performed. *npM1*, having completed 5 more procedures since the last milestone, finishes with a slightly lower estimated mean procedure time 24:42 with narrower 95% CI [21:01, 28:27]. *npB1* had finished 13

procedures more than at the time of previous milestone and ends with a lower estimated mean time 31:21 with 95%CI [25:27, 41:32].

2.6.3 Errors in manual entry of blood glucose measurements

Out of 227 patients assessed, 14 patients had an erroneous value of glucose level entered. The median error was 9.55 mg/dL with IQR 35.75 mg/dL (0.53 mmol/L with IQR 1.99 mmol/L).

Usefulness of point-of-care management recommendations for ASHAs Table 2.6 quantifies the extent to which the built-in management guidelines for CVD were used by the ASHAs through the number of clicks recorded in each management section (outlined previously in Table 2.1) for all procedures completed.

Table 2.6 Statistics for how often the management recommendations in the CDSS were actually used by the ASHAs. The usage of the risk projection meter was much lower in comparison to the other sections for the three ASHAs. Also only 81% participants screened by ASHA *npL1* were told about their next visit for follow-ups while in contrast, ASHA *npB1* disseminated the information to 93% of her participants.

ASHA	**Risk projection meter to communicate CVD risk association between risk factors**	**Recommendations** (lifestyle, smoking, nutrition)	**Next visit** (physician referral, CVD risk/diabetes screening)
npB1	79% (22/28)	96% (27/28)	93% (26/28)
npM1	75% (15/20)	85% (17/20)	90% (18/20)
npL1	81% (13/16)	81% (13/16)	81% (13/16)

2.6.4 CVD referrals

An important component of the mobile-based CDSS is the referral indication (as mentioned in Section 2.2.3). Figure 2.10 shows the major reasons for referral. Out of 227 participants screened by the ASHAs, the CDSS identified 57% (n=128) for referral to a physician either for high CVD risk (n=88) or IFG (n=40). When participants were followed up a month after

they had been referred, it was found that out of the 57% who were referred, only 30% of them visited a physician. Amongst those who visited the physician, 57% went to a PHC while 43% saw private physicians.

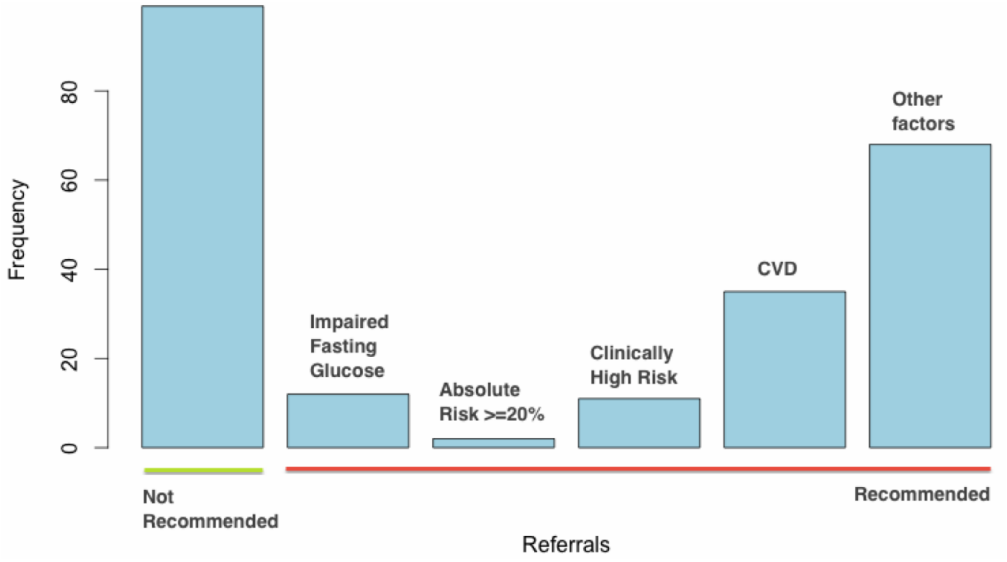


Figure 2.10 Rationale behind physician referrals formulated by the CDSS for ASHAs (N_{asha}=227)

2.6.5 Usability

The end-users of the application were asked to complete a questionnaire at the end of every risk assessment procedure. In over 72% of the screening procedures performed, the mobile application was found easy to use for that particular procedure. Users concurred similarly on the usefulness of the graphic bar that visualised risk scores in communicating the meaning of absolute CVD risk reduction to the community participants. No user gave the application a rating below 3 on a scale of 4, with 4 being most useful and 1 being least useful. In less than 2% of the procedures performed, the ASHAs recorded difficulties with collection of risk factors (such as blood pressure, height and weight, blood glucose).

2.7 Discussion

The CDSS was designed and developed using the agile development methodology with the close engagement of the ASHAs and physicians. The process was iterated until ease-of-use was confirmed by the end-users and this iteration became the final version of the CDSS. A number of key features were added through this process. The use and integration of open-source telemedicine platforms such as OpenMRS and SanaMobile were intended to increase interoperability. This is important because in the area of mHealth application development, there exist silos of numerous non-interoperable mHealth applications, increasing the risk of duplication in effort and fragmentation of purpose. The risk profiles of screened participants in the three villages in Andhra Pradesh had a bimodal risk distribution, where the participants were mostly either low risk or high risk. The number of participants at high risk of CVD was almost one third of the screened sample. A decrease in median completion time was observed as more procedures were performed. This indicated the ASHAs increased familiarity with the process and generates evidence for variability in usage. The median procedure time for ASHAs was 21 minutes for a complete CVD risk assessment with Step-3 (risk factor acquisition) taking the maximum time (over 10 minutes although with large variations between ASHAs). Acquisition of three BP measurements, on its own, took approximately 7 minutes. This could be useful to assess the maximum number of community participants that can be screened by an ASHA given her usual commitments to antenatal care.

It is interesting to compare the usage patterns of the three ASHAs (*npB1*, *npM1*, and *npL1*) from the four parameters that we have analysed - total number of risk assessments, median completion time and 95% CI, use of Bluetooth functionality, and rate of disseminating management recommendations to participants. *npB1* may be classified as one of the better performing ASHAs in the pilot study as, firstly, she performed the highest number of

risk assessments amongst all ASHAs. Secondly, she had the highest usage rate of Bluetooth functionality and lowest mean BP acquisition time. Thirdly, she had the highest rate of disseminating management recommendations although her mean procedure time and 95% CI was fairly average compared to other ASHAs. *npM1* showed more consistency than any other ASHA in all aspects of her performance. The number of assessments she completed was exactly the same as the approximate target of 20 procedures prescribed for all ASHAs at the start of the pilot. With regard to the automated Bluetooth feature, the usage rate was almost the same before the ASHA's 10^{th} procedure (50% usage rate) and after her 10^{th} procedure (44% usage rate). With the lowest estimated mean time and smallest 95% CI as well as an approximately average rate of disseminating management recommendations, ASHA *npM1* was the most consistent performer. *npL1* performed the fewest assessments, had the highest estimated mean time with a wide 95% CI, low dissemination rate of management recommendations, and the lowest usage rate of the automated Bluetooth feature. However, with regard to Bluetooth usage, we observe that even though only a single attempt (or 10%) was made before her 10^{th} procedure, her usage rate increased to 80% thereafter. Though this may indicate a longer learning curve, ASHA *npL1* may be identified as one that would need additional attention or performance monitoring. By recording user interactions with the mobile application, we could thus gauge the end-user's behaviour and effectiveness of the tool's features such as the rate of dissemination of point-of-care CVD management recommendations.

The introduction of automated Bluetooth BP measurements in this study found low adoption initially. However, we observed that towards the end of the study, the ASHAs recognised the ease of use and utility of this feature, as was observed through qualitative interviews (not reported in this chapter). ASHAs also recorded difficulties with operation of the BP monitor. Given the current overall adoption rate of 55%, a less sophisticated BP monitor with

faster acquisition time may increase adoption and optimise the time taken for Step 3. The errors reported from manual data entry during blood glucose measurement demonstrate that transcription and transposition errors were likely and the erroneous values had a wide range. With regard to POC decision support, communication through the risk meter found the lowest adoption (Table 2.6) and this is possibly due to difficulties in the ASHAs understanding of the risk projection feature. The overall trends in user interactions were derived from our analytics framework in the CDSS. This was beneficial to envisage learnability [52], an indicator of how easily the ASHAs performed tasks as well as efficiency [53], an analysis of the ASHAs ability to swiftly perform the task they learnt. The understanding of these parameters can aid in better design (for example, by evaluating features widely used versus those unused or difficult to use) and unravel user performance over time so that adequate support can be provided to those ASHAs requiring technical assistance (for instance if the ASHA's efficiency declines, as her estimate of mean procedure time increases successively).

By involving the end-users iteratively as well as acquiring information through data analytics, we minimise social desirability bias [54][55], an important issue in acceptance testing in global health. Another key aspect of our study is the design thinking process that has cohesively involved the local community participants and end-users, and therefore differs from global health projects which 'push' solutions onto the community. The in-depth design strategies resulted in a tool that could be conveniently integrated within the workflow of the ASHAs, as reflected by recorded consensus on ease-of-use, usefulness, and in increased speed of assessments over time.

A major limitation in this pilot study was the limited sample size; however, this was not designed to be representative. Our study also did not conduct follow-up of individual participants through referral pathways and hence was not designed to assess longitudinal

impact. Infrastructural barriers observed in our pilot study included power constraints as power supply was limited to fixed times of the day, and so the low-cost tablets had to be charged frequently.

2.8 Conclusion

This chapter has described the development an evidence-based CVD risk prediction and management tool suitable for rural India, designed with proper engagement of the end-users, health care providers, and local communities. Field evaluation established the platform to be feasible and widely accepted by all stakeholders involved. A quantitative exploration of successful user engagement could be gauged through our analytics platform via detailed usage patterns. With over a third of screened participants being high risk, and gender-specific differences in managing the level of risk factors such as BP, there is a need to demonstrate longitudinal impact of the mHealth platform so that it could contribute to improved CVD detection in high-risk low-resource settings. In the next chapter, the CDSS tool will be utilised to collect large-scale risk factor data from 54 villages in Andhra Pradesh as a prelude to RCT evaluation.

Chapter 3

SMARThealth baseline study: large-scale data collection using an mHealth system

3.1 Introduction

The last chapter presented the design and development of the *SMARThealth* CDSS for CVD risk assessment and management. A pilot test with 292 participants from rural Andhra Pradesh demonstrated preliminary feasibility and acceptability. The advantage of having an electronic CDSS in comparison to paper-based systems is the ability to scale-up, which is essential in countries with a large population such as India. A previous study from the clinical collaborators in the SMARThealth-India programme, the George Institute for Global Health, collected mortality data from the Godavari district between 2003 and 2004 [56]. The authors, Joshi et al. [56] , confirmed that CVD was the leading cause of mortality, and showed that it was responsible for at least 32% of all deaths. The mortality registries of 53 villages that the authors studied revealed half of all CVD related deaths were of people below

70 years of age.

The focus of this chapter is on the acquisition of large-scale data on CVD risk factors using *SMARThealth*. We present a mobile-based data collection and screening of participants across 54 villages that were broadly representative of the West Godhavari district, Andhra Pradesh (region highlighted in Figure 3.1). This study was performed to record baseline risk factors in the population (and hence is called the 'baseline study') as a precursor to a Randomised Controlled Trial (RCT) that aims to evaluate the effectiveness of our mHealth CDSS.

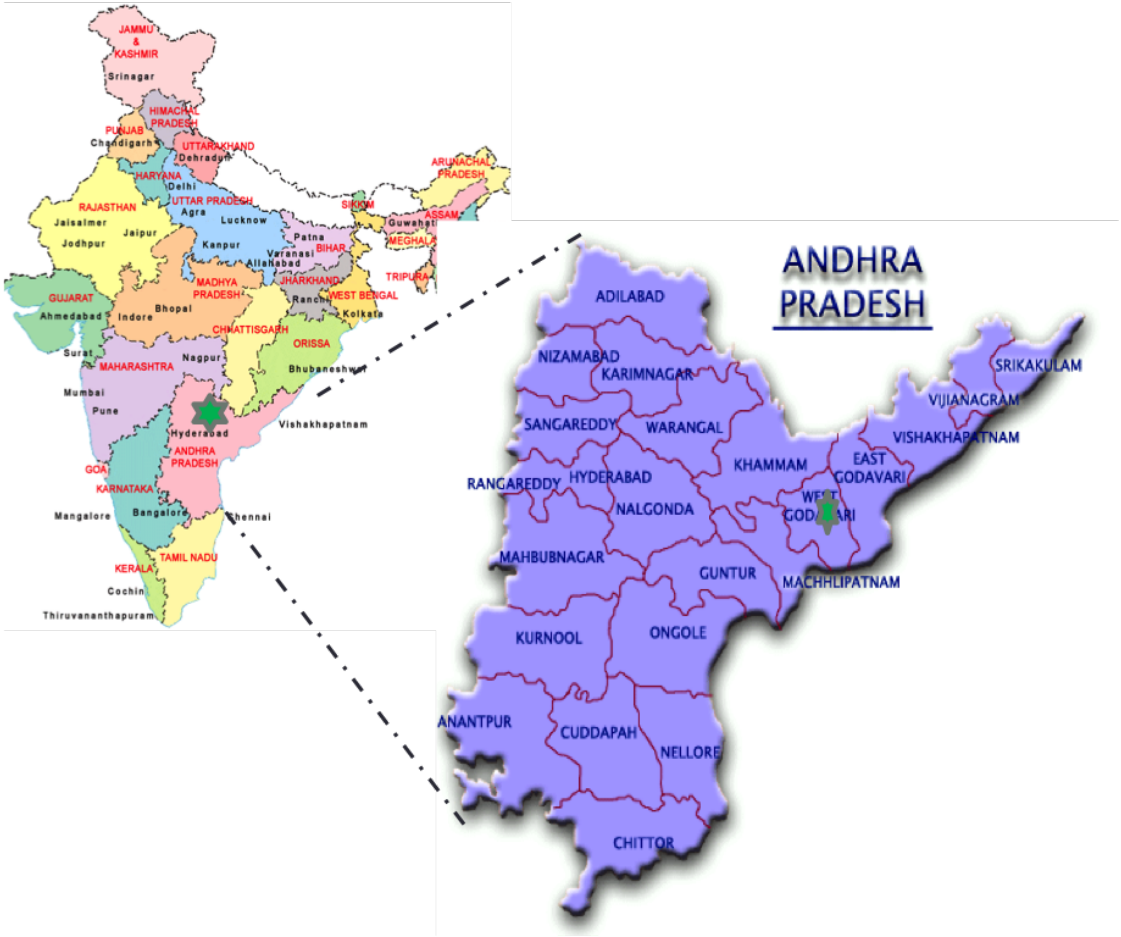


Figure 3.1 Site of baseline data collection in rural India.

3.2 Methods

3.2.1 Data collection process

Fifty-four villages around the major town in West Godhavari (WG) district, Bhimavaram, that were broadly representative of the region were identified. Selection of villages was performed by the George Institute of Global Health, which has an established history of working with health providers and non-governmental organisations in the WG district. The major selection criterion for villages. however, was that the nearby Primary Health Centre (PHC) had to have at least one physician offering regular services. This baseline data collection was conducted between February and May 2014. Based on the 2011 government census list on the WG district of Andhra Pradesh, the inhabitants residing in each household of the 54 villages were identified. Subsequently, trained interviewers performed paper-based verification of the census list (hence called 'listing data'). Specifically, residents from all households were interviewed to reconfirm age, gender, and number of family members from the listing data, and incorrect entries were modified. Focus was given to residents between 35 and 42 years of age, who confirmed the correctness of their data so that the risk of missing eligible participants was minimised. This was necessary as eligibility for the baseline study required that participants were aged 40 years or above. This process constituted Phase 1 of data collection.

Phase 2 utilised the *SMARThealth* mobile application that was used for pilot data collection as described in Chapter 2, with minimal modifications. The technical details of the modifications in each step of data collection are presented in the next Section 3.2.2. The listing data from Phase 1 was packaged into the *SMARThealth* application in a distributed form (the listing data specific to the location or village only was loaded into the tablet used in that location/village). This enabled trained interviewers to perform localised door-to-door

data collection as well as assess risk since they had accurate guidance on eligible participants in each household. The entire data collection process is summarised in Table 3.1.

3.2.2 mHealth infrastructure

In phase 2, trained interviewers were given a Samsung Galaxy 3 tablet with the *SMARThealth* application. Although the application used the pilot version, minimal modifications were performed to enable the recording of additional parameters once a participant was found to be high risk, as well as to assist the monitoring of the data collection process. Figure 3.2 describes the various steps of the data collection using the *SMARThealth* application on the Samsung tablet. The first 3 steps of data collection (including patient registration/demographics, medical history, and risk factor assessment) were largely the same as for the pilot application. The only exception, in Step 1, was the inclusion of questions on literacy status, marital status, and occupation. For Step 3, three blood pressure measurements were acquired for each participant at 2-minute intervals with a sphygmomanometer (Model UA-767PBT-C40, A&D Medical, Tokyo, Japan) that uses the oscillometric method. Care was taken to ensure participants were rested for a minimum of 5 minutes prior to the three BP measurements. Capillary blood was sampled using the finger-prick method for blood glucose estimation using a point-of-care blood glucometer (Abbott FreeStyle Optium, Alameda, California, USA). Participants were asked for when they had last eaten, and those who had not consumed food for at least 6 hours prior to measurement were considered to be fasting.

As observed in the last chapter, Step 3 took the longest in terms of data acquisition. Therefore to achieve a tradeoff between time for data collection and the requirement to be more efficient to enable scale-up, certain parameters were collected only for those at high CVD risk based on recommendations of the collaborating clinicians. Height and weight in

Table 3.1 Table explaining the methodology of SMARThealth baseline study.

	Phase 1 Household listing	**Phase 2** CVD risk screening (Baseline)
Process	Based on the 2011 government census for the WG district of Andhra Pradesh, the inhabitants from each household were identified. Subsequently, trained interviewers performed paper-based verification of census data namely age, gender, and number of inhabitants per household across the 54 villages in WG district.	The *SMARThealth* mobile application incorporated listing data obtained from Phase 1. Interviewers used the mobile application to survey eligible inhabitants (age≥40) and identify high-risk participants. To achieve maximum coverage and reduce the risk of gathering incomplete data, three additional visits were attempted if a participant was unable to be interviewed initially. We achieved this through synchronisation of crucial data between the mobile tablet and server.
Time-period of data collection	December 2013- February 2014	February 2014- May 2014
Data collection tool	Paper based	Mobile-based (SH Baseline app)
Target Population	Inhabitants in rural Andhra Pradesh.	Eligible individuals from Phase 1 who gave written consent for participation.
Decision support		Before screening, borderline aged participants between 38-40 years were re-checked for the accuracy of their age. If a participant was found to be high risk, an extended baseline questionnaire was administered. For ethical reasons, participants with SBP≥180 or DBP≥100 were referred to a physician.
Data Description (Variables)	Household Address, village, total member count, eligible member count, first name, last name, gender, age	Phase 1 parameters and demographics, medical history, treatment history, risk factors (blood pressure, blood glucose). For high-risk participants, medication use, quality of life (EQ-5D) [57], physical activity (IPAQ), well-being index (WHO) were recorded.

particular, were collected for only high-risk participants. After Step 3, additional parameters were collected if the participant was found to be at high risk. The additional input variables are listed in Table 3.1.

Physical activity can be evaluated with different methods. However, in order to follow a standardised procedure, physical activity was measured using the International Physical Activity Questionnaire (IPAQ) [58]. Similarly, standardised measures of health-related quality of life and well-being were measured using EQ-5D [57] and the WHO well-being index [59], respectively. The questionnaires (Physical activity, quality-of-life EQ-5D, WHO Well being index) were translated into the local language Telugu, and back-translated to reduce translation errors. Permissions were obtained from the EuroQol group to translate the EQ-5D quality of life assessment in Telugu, as well as in the form of a mobile application.

A screenshot from EQ-5D is shown in the 'Additional parameters' screen in Figure 3.2. High quality data was achieved through three mechanisms, all of which were tested in the pilot study. Firstly, each input variable was checked for the data range. Secondly, completion of all input variables was mandatory before the user could proceed to the next step. Thirdly, a convenient mechanism to manually edit input variables at any step of the data collection process was provided. For instance, the 'Step 2' screen in Figure 3.2 exemplifies data checks for quality improvement.

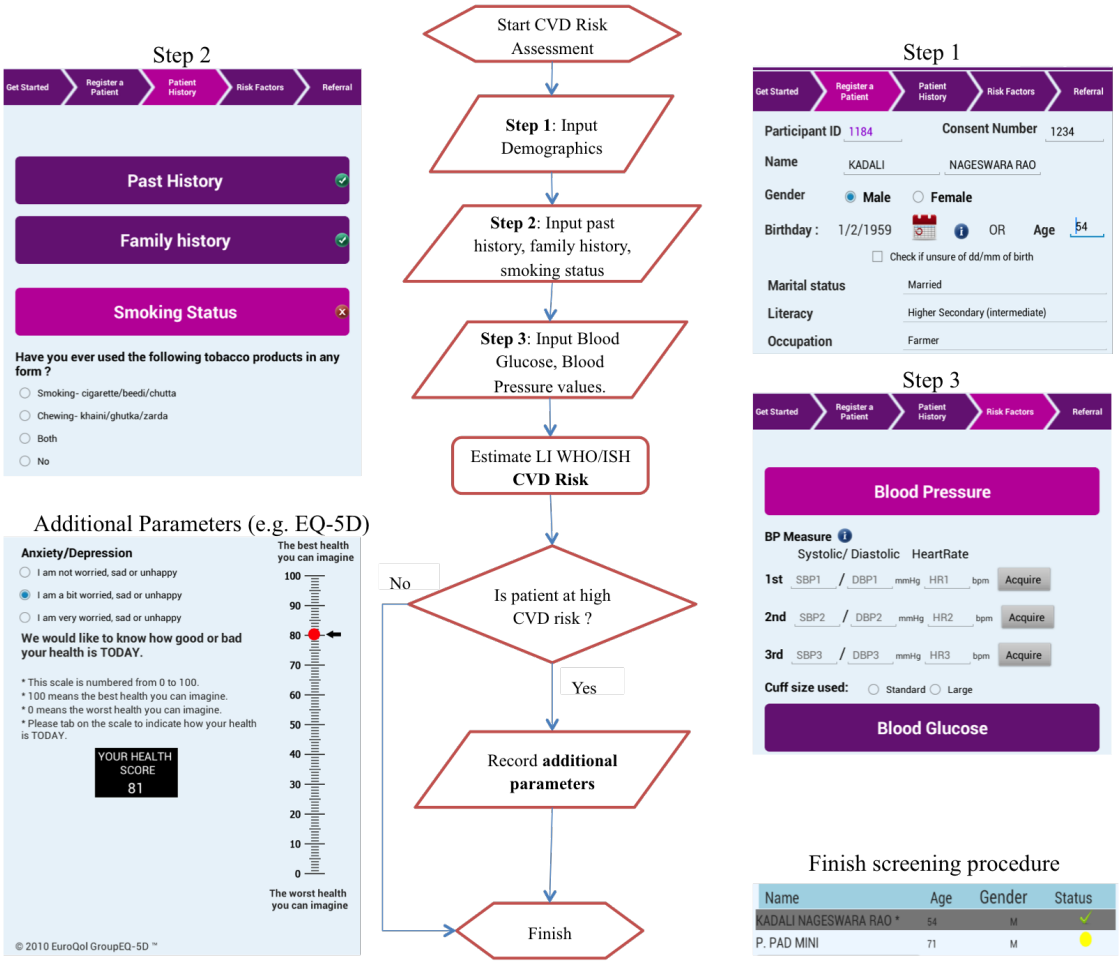


Figure 3.2 Data collection process on the mobile tablet with the *SMARThealth* application. Screenshots from Steps 1-3 are illustrated. Participants who were at high risk of CVD had additional data recorded, such as the EQ-5D (screenshot shown). Once data collection was finished for the participant, their 'Status' was updated, which was then synchronised with their information on the server. The participant name and data shown here are for illustrative purposes only.

The salient features of the *SMARThealth* application are discussed below.

1. **Participant and Household ID generation**: An identifier (ID) for each participant was generated after phase 1 of the data collection process. Sequential ID numbers were given to each PHC, village, and household. Participant ID was designed to be eight digits long, and was a concatenated numeric code of PHC ID (contributed 2 digits),

village ID (contributed 2 digits), household ID (contributed 4 digits), and participant number in household (contributed 2 digits).

2. **Handling areas with weak internet coverage**: Lessons from the pilot study presented in the last chapter showed the lack of robust 3G networks in many regions of the West Godhavari district. To handle weak or fluctuating signal strength, two fixes were incorporated into the *SMARThealth* application:

   (a) Data on the client-side tablet that failed to be transmitted to the server was *queued* on a separate SQLite database. This queue was routinely *cleared* depending on the speed and strength of the available connection. Connectivity was checked (a) at specific intervals (07:00 hours, 12:00 hours, 21:00 hours) during the day and (b) whenever new data was received or transmitted by the mobile-tablet. When the signal strength was stable for over 30 seconds, the process of re-sending queued data was initiated. We performed preliminary A/B testing, which involved implementing the modifications (*queueing* and *clearing*) on one half of all deployed Samsung tablets whilst maintaining a vanilla version (or status quo) on the other half. This procedure revealed that the *queueing* and *clearing* mechanisms substantially reduced data attrition rates.

   (b) The 2G/EDGE network was used for all data transmission and retrieval. This avoided automatic switching between 2G and 3G speeds on account of fluctuation.

3. **Increasing coverage of data collection**: In the process of collecting data from villages, we observed instances where information previously registered from Phase 1 was incorrect or changed with time, and scenarios where data collection could not be fully completed for a participant. To mitigate these issues and streamline the process, two

levels of additional detail were captured. These included the Household Disposition Code (HDC) and Respondent Disposition Code (RDC).

(a) **Household Disposition Code (HDC)**: If information on a previously registered household changed, the interviewers could include additional information. Four scenarios were provided:

- If the household has moved away permanently from the village and new people are living in that house instead → [option to enter details of eligible individuals in the new household]
- If the listing has one household but two families are currently residing → [option to enter details of eligible individuals in both households, if not listed]
- New household → [option to enter details of eligible individuals in the new household]
- Household refuses to participate → [remove household from study]

(b) **Respondent Disposition Code (RDC)**: If an eligible participant was unable to be screened and/or if data could not be collected in full, the follow scenarios were provided to the interviewer:

- Respondent moved away permanently → [name is greyed out and data is removed]
- Respondent not at home during the visit → [option to click on this respondent on next visit]
- Respondent refuses to participate → [option to click on this respondent on next visit]
- Interview not completed → [option to click on this respondent on next visit]

- Incorrect entry of a respondent who is actually not eligible → [name is greyed out and data is removed]
- Respondent is dead → [name is greyed out and data is removed]
- Respondent is incapacitated → [name is greyed out and data is removed]

Both HDC and RDC were recorded prior to Step 1 of the data collection process and provided useful meta-data that accurately captured the number of participants covered in each household within every village. This highlights one of the many advantages of using an electronic tool. It can capture granular information, which, in our case, is key to maximising the coverage of data collection from the eligible inhabitants in each village.

4. **Server side electronic health record**: Similar to the pilot study architecture in Chapter 2, version 1.9 of the Open Medical Record System (OpenMRS) [60] was installed to create health records for all participants. Additionally, we created a web-application that provided an interface with OpenMRS. By following a similar database schema, our web-application was made interoperable and not built as a siloed system. A key improvement of the mHealth system architecture was the synchronisation of data and meta-data (including RDC and HDC) from every mobile tablet with OpenMRS. The eight digit participant ID was stored as an 'attribute' [61] which allowed faster retrieval of associated data and synchronisation. Our mHealth system thus achieved two key features in the baseline study:

   (a) **Data monitoring**: Every set of interviewers reported to a project manager who monitored the progress of the data collection process. Most project managers demonstrated familiarity and understanding of the popular spreadsheet application Microsoft Excel. However, they were not comfortable running queries with existing data monitoring modules that OpenMRS provided such as the Reporting

module [62]. Furthermore, running multiple queries on OpenMRS 1.9 from different end-users (project mangers in this case), resulted in server down time which had to be solved by restarting the server. Therefore, our web-application was vital in facilitating 4 key elements required to run a successful baseline study, as described in Figure 3.3. This facilitated 4 key elements required to run a successful baseline study:

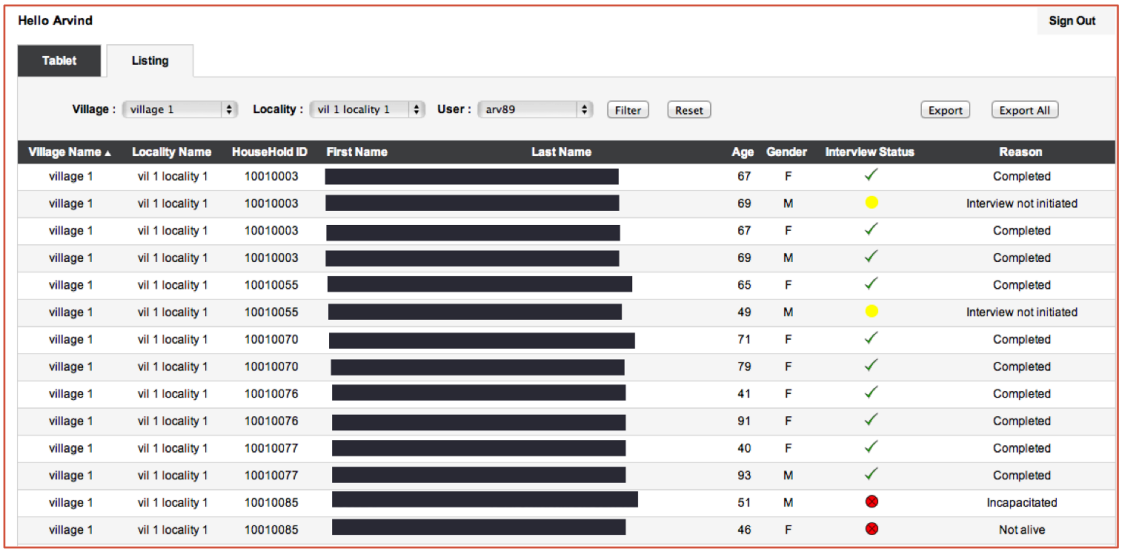
i. *Utilising existing workforce*: convenient and easy-to-use interface for project supervisors and managers to monitor field data collection; the web-application offered an interface similar to Microsoft Excel which was highly familiar to project supervisors.

ii. *Enabling regular monitoring*: The web-application provided a simple tool for exporting data. It also included options for filtering using ad-hoc parameters such as village and user ID.

iii. *Interoperability*: The web-application leveraged on the existing OpenMRS database schema and can thus be extended for longitudinal evaluation.

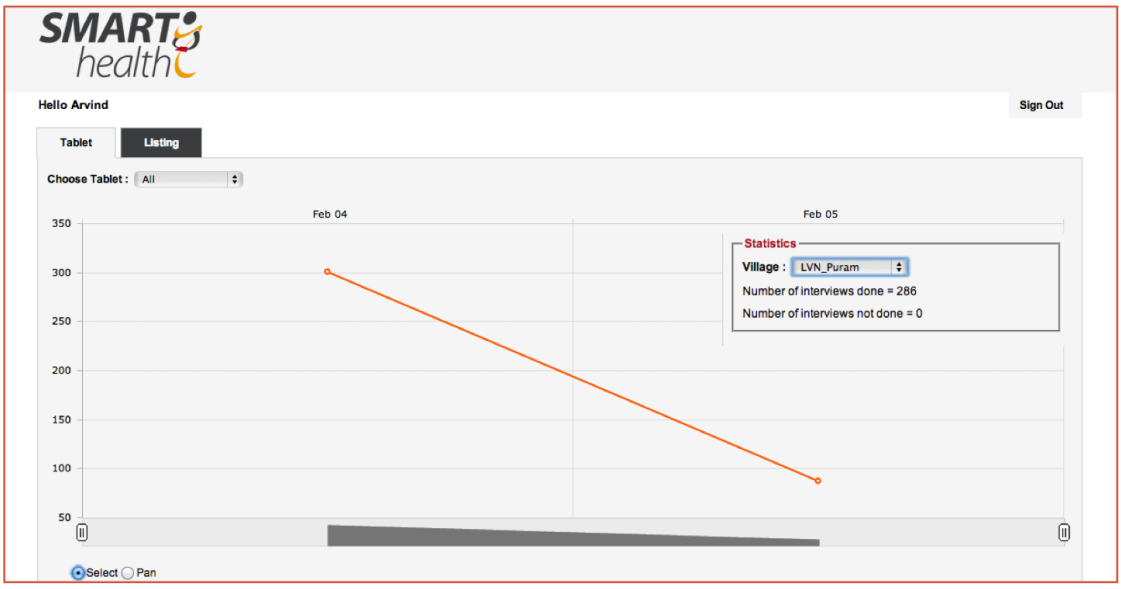
iv. *Visualisation*: Graphic description that quantified the progress of the data collection process.

(b) **Identifying and planning re-visits**: Each eligible participant whose data could not be recorded on the first visit were revisited up to two times. This was achieved by enabling the accurate identification of those participants or households missed during the initial visit for each of the 54 villages.



Village Name	Locality Name	HouseHold ID	First Name	Last Name	Age	Gender	Interview Status	Reason
village 1	vil 1 locality 1	10010003	[redacted]		67	F	✓	Completed
village 1	vil 1 locality 1	10010003	[redacted]		69	M	●	Interview not initiated
village 1	vil 1 locality 1	10010003	[redacted]		67	F	✓	Completed
village 1	vil 1 locality 1	10010003	[redacted]		69	M	✓	Completed
village 1	vil 1 locality 1	10010055	[redacted]		65	F	✓	Completed
village 1	vil 1 locality 1	10010055	[redacted]		49	M	●	Interview not initiated
village 1	vil 1 locality 1	10010070	[redacted]		71	F	✓	Completed
village 1	vil 1 locality 1	10010070	[redacted]		79	F	✓	Completed
village 1	vil 1 locality 1	10010076	[redacted]		41	F	✓	Completed
village 1	vil 1 locality 1	10010076	[redacted]		91	F	✓	Completed
village 1	vil 1 locality 1	10010077	[redacted]		40	F	✓	Completed
village 1	vil 1 locality 1	10010077	[redacted]		93	M	✓	Completed
village 1	vil 1 locality 1	10010085	[redacted]		51	M	✗	Incapacitated
village 1	vil 1 locality 1	10010085	[redacted]		46	F	✗	Not alive

(a) Web application for monitoring and exporting data collected from 54 villages, 18 PHCs in Andhra Pradesh. Names have been masked for de-identification. The last two columns on the right ('Interview Status' and 'Reason') provided project managers with remote information on the response rate and thereby improved the efficiency of the large-scale data collection process.



(b) Web application for visualising number of participants screened over the duration of the data collection process. The number of participants assessed for CVD risk in the village 'LVN Puram' on February 4^{th} and 5^{th} are marked. The line joining the points shows a visible decrease in the number of screened participants thereby graphically depicting the efficiency of data collection per day (or week/month as can be adjusted by the Pan function shown at the bottom of the image).

Figure 3.3 Server-side web application built as an extension to OpenMRS medical record system.

3.2.3 Clinical definitions

We defined a case of **suspected diabetes** according to the American Diabetes Association 2010 guidelines [63]. A person was suspected to be diabetic if his or her random capillary glucose value was $\geq$200 mg/dL (11.1 mmol/L) or if their fasting plasma glucose was $\geq$126 mg/dL (7.0 mmol/L) or that person had previously been diagnosed with diabetes (but not gestational diabetes) by a physician. The definition of **hypertension** or high blood pressure was as specified by the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC7) guidelines [64]. The last two BP readings were used to compute a mean BP measurement. Hypertension was defined to be: mean systolic blood pressure (SBP) $\geq$140 mmHg and/or mean diastolic blood pressure (DBP) $\geq$90 mmHg, or previous diagnosis of hypertension by a physician.

With respect to the use of tobacco, a '**current smoker**' was a participant who had smoked in the past month while a '**current chewer**' would have chewed tobacco products in the last month. Only current smokers are considered in this chapter due to ambiguity in quantifying 'current chewers'.

We define **CVD risk** based on the WHO and Indian National Programme for for Prevention and Control of Diabetes, Cardiovascular Diseases, and Stroke (NPCDCS) guidelines [1]. **Low risk** was defined as a 10-year CVD risk less than 10%. **Intermediate risk** was defined as a 10-year CVD risk of between 10 and <20% or 20 to <30% with SBP<140 mmHg. **High risk** was defined as the presence of one or more of the following: Past history of CVD (including myocardial infarction/stroke/peripheral vascular disease), 10-year CVD risk $\geq$ 30%, or 10-year CVD risk of 20-29% and SBP>140 mmHg. **Established CVD** was based on past history of myocardial infarction/stroke, peripheral vascular disease, or angina. A **clinically high risk** condition was SBP $\geq$ 160 mmHg or DBP $\geq$ 100 mmHg.

3.2.4 Multivariate data visualisation

In this section, the aim is to graphically present a low-dimensional representation of the key risk factors collected from the large number of participants in the West Godhavari district of Andhra Pradesh in rural India. This visualisation is a first line of exploratory data analysis and can help to discern any underlying patterns in the data structure. In public health, collecting data from more individuals is usually favoured, given that it will increase representativeness of the dataset for the target region. However, an associated challenge during analysis is the visualisation of multi-dimensional data with large samples. A popular approach is dimensionality reduction, where the objective is to preserve, as much as possible, the structure of high dimensional data in a low dimensional space. Low dimensional embedding involves condensing high dimensional distances between features into equivalent distances in two or three dimensions.

A plethora of visualisation techniques exist, and reviews can be found in De Oliveira et al. [65] and Heer et al. [66]. Classical Principal Component Analysis (PCA) [67] works well when the data is linearly separable but may require modifications to deal with non-linear manifolds in underlying data structure. For instance, in the case of a synthetic dataset that resembles a 'swiss-roll', points which are nearby in Euclidean space may not be so close when considering the entire manifold of the data structure [68].

Non-linear techniques for dimensionality reduction can be broadly categorised into two types [69]: (a) techniques that aim to maintain global properties of data in the low-dimensional space such as Multi-Dimensional Scaling [70]; (b) techniques that aim to maintain the local properties of data in the low-dimensional space such as Local Linear Embedding [71]. Although non-linear techniques have shown good performance on synthetic datasets, they have not always outperformed linear techniques on natural datasets [69]. Van

der Maaten [69] performed a comparative review of linear and non-linear dimensionality reduction techniques, and reported two findings: (a) techniques aiming to preserve local properties are largely limited by the dataset dimensionality and (b) the extent to which a technique retains local and/or global properties is more important than the type of property itself. Van der Maaten and Hinton [72] criticise other classical techniques including Sammon mapping [73], Isomap [74], and Stochastic Neighbour Embedding [75] for their inadequate retention of the local and global structures of the input data within a single map.

An unsupervised dimensionality reduction technique that is capable of handling large, high-dimensional datasets and achieves low dimensional embedding is tSNE (t-Stochastic Neighbour Embedding) [76]. tSNE builds on Stochastic Neighbour Embedding (SNE), originally proposed by Hinton et al. [75]. In contrast to the traditional techniques mentioned so far, tSNE captures both the global and local structure of high-dimensional data. Van der Maaten [76] claimed two advantages of tSNE over PCA. Firstly, the linear nature of PCA is a limitation as most medical data has non-linear dependencies. Secondly, PCA aims to preserve large pairwise distances in the low dimensional space whilst not laying enough emphasis on preserving local structure.

We first describe the methodology behind SNE and subsequently discuss tSNE along with its advantages and variations. SNE transforms Euclidean distance between high-dimensional data to conditional probabilities that represent similarity between data points. If we have a high dimensional input space $D = x_1, x_2 ... x_N$, where N is the number of data samples, the similarity of a datapoint x_j to x_i can be given by the conditional probability $p_{j|i}$. This can be understood as follows: x_j would be selected as a neighbour to x_i if the probability of selecting a neighbour was proportional to the width of a Gaussian kernel σ_i centered around

the datapoint x_i. Naturally, nearby points would have high $p_{j|i}$ while farther points have largely decreasing $p_{j|i}$. We represent this mathematically as:

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{3.1}$$

Similarly, $q_{j|i}$ is the low dimensional counterpart of $p_{j|i}$ representing conditional probabilities. SNE uses a Gaussian kernel to compute $q_{j|i}$, similar to Equation 3.1 albeit with a fixed variance of $\frac{1}{\sqrt{2}}$. The fixed variance allows for mathematical convenience (since $2\sigma_i^2$ becomes 1). Both $p_{i|i}$ and $q_{i|i}$ are set to 0 since only pairwise distances are of interest.

$$q_{j|i} = \frac{exp(-\|y_i - y_j\|^2)}{\sum_{k\neq i} exp(-\|y_i - y_k\|^2)} \tag{3.2}$$

We now have two conditional probabilities representing similarities across both the input (or high dimensional) space and the low dimensional space. SNE is motivated by the premise that if the similarity between two points x_i and x_j in the high dimensional space are correctly modeled by two corresponding points in the low dimensional space y_i and y_j, then $p_{j|i}$ and $q_{j|i}$ will be equal. This implies that minimising the mismatch between $p_{j|i}$ and $q_{j|i}$ can help uncover the low dimensional representation. SNE uses the Kullback–Leibler (KL) divergence which measures the difference in two distributions P and Q, and is denoted by $D_{KL}(PQ)$. The sum of KL divergences over all data points represents the cost function (C) to be minimised in order to obtain the low dimensional representation.

$$\text{argmin} C = \sum_i D_{KL}(P_i Q_i) = \sum_i \sum_j p_{j|i} log(\frac{p_{j|i}}{q_{j|i}}) \tag{3.3}$$

where P_i represents the conditional probability (or similarities) of data point x_i with all other data points. C can be minimised by gradient descent.

The tSNE technique proposed by Van der Maaten [72] has two major differences with respect to the SNE method described above. Firstly, 'symmetric' SNE is used, which means that instead of minimising the sum of the KL divergences between the conditional probabilities $p_{j|i}$ and $q_{j|i}$, a single KL divergence minimisation is performed between the joint probability distribution P in the high-dimensional space and the joint probability distribution Q, in the low dimensional space. This implies that $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$. 'Symmetric' SNE provides a more convenient and faster computation of the gradient of C. Secondly, tSNE utilises a Student t-distribution with one degree of freedom instead of a Gaussian kernel because the heavy tail allows more space for modelling small pairwise distances in the low-dimension [72]. This circumvents the 'crowding problem' observed in SNE, where accurate low dimensional representation of nearby points is difficult since points farther away have to be located at much greater distances. The crowding problem can also be observed in other dimensionality reduction techniques such as Sammon mapping [72].

A key challenge in tSNE is performing arithmetic calculations using joint distributions that raise the complexity to $\mathcal{O}(N^2)$. This quadratically scales the computational demand and is inefficient for datasets with more than a few thousand samples [77]. To overcome this obstacle, a faster implementation of tSNE was proposed by Van der Maaten [77] that implemented a variation of the Barnes-Hut algorithm for reducing computational complexity. This stated two further modifications to the standard tSNE algorithm. The first modification involves performing a sparse approximation to calculate similarities in the high-dimensional space. The second modification requires the incorporation of interactions between a group of points as opposed to pairs of points, and helps in deriving an approximation of the gradient of C.

In this thesis, the MATLAB implementation the Barnes-Hut variant of tSNE available from van der Maaten [77] is used. This leads to faster performance with our dataset which includes 54 villages with well over ten-thousand data points owing to reduced computational complexity, ($\mathscr{O}(N\log_{10}N)$ instead of $\mathscr{O}(N^2)$), and lesser memory requirements ($\mathscr{O}(N)$).

In order to provide a more intuitive interpretation of the clusters formed by tSNE and to serve as a basis for comparison of the underlying data structure, the Neuroscale visualisation technique [78] was also used. Neuroscale utilises a Radial Basis Function (RBF) neural network to map m feature vectors with N data points to a low-dimensional space. Consider two points x_i and x_j in the high-dimensional space, and two corresponding points y_i and y_j in the low-dimensional space, the RBF neural network works to minimise $\sum_i^N \sum_{j>i}^N (d_{ij}^* - d_{ij})^2$, where d_{ij}^* and d_{ij} are the inter-point Euclidean distances in the high-dimensional space and low-dimensional space, respectively. The application of an RBF kernel applies a non-linear transformation, given by $\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i; \mathbf{W})$ where $\mathbf{W}$ is a vector of the weights. The architecture of the Neuroscale neural network is shown in Figure 3.4.

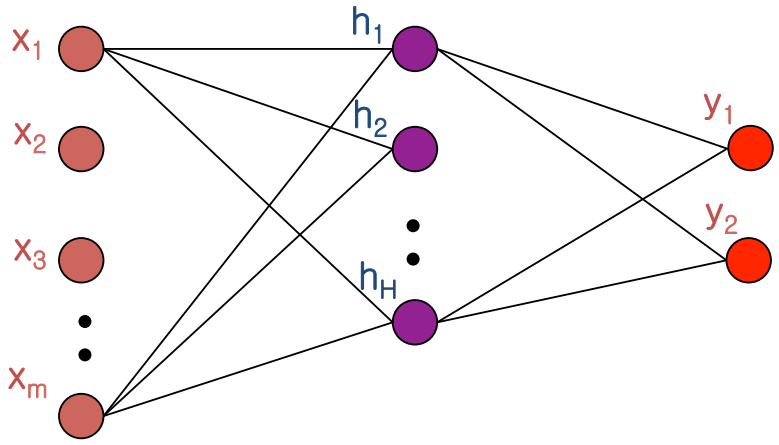


Figure 3.4 The Neuroscale visualisation technique that maps m input features given by $\{x_1, x_2, x_3, \ldots, x_m\}$ to hidden layer of H nodes. The hidden layer maps to a low-dimensional space given by the output layer of 2 nodes (hence to a 2 dimensional representation).

The neural network consists of three layers namely the input layer M, the hidden layer H, and the output layer P. The weights $\mathbf{W}$ determine the mapping between the hidden layer and the output layer, and therefore determine the location of points in the low-dimensional space. The advantage of Neuroscale is that once the weights have been learnt, much larger datasets (not included whilst training the neural network) can be projected.

The first 8 villages, comprising 10,000 participants free of previous incidence of cardiovascular diseases were visualised using the techniques described above, namely tSNE and Neuroscale. Prior to visualisation, data was normalised using the z-score normalisation given by

$$\mathbf{x}_j^* = \frac{\mathbf{x}_j - \mu_j}{\sigma_j} \tag{3.4}$$

where $\mathbf{x}_j^*$ is the normalised value of $\mathbf{x}_j$, μ_j and σ_j are the mean and standard deviation for the j^{th} feature, respectively.

Performing visualisation with mixed data types (involving a combination of continuous, binary, categorical, or ordinal variables) is not straightforward. In our dataset, after normalisation, binary variables (such as gender) often have higher densities at the extremes of the range of the variable in comparison to the extremes of the range of continuous variables (such as age). Therefore the computation of inter-point distances can be biased depending on the type of the variable. For the task of visualisation in this chapter, however, continuous variables namely age, SBP, DBP, glucose levels, recorded for participants across 54 villages in Andhra Pradesh are used. The visualisation is performed separately for males and females, smokers and non-smokers, as well as for participants were and were not currently undertaking treatment for hypertension. The resultant data points on the low-dimensional space are coloured with the WHO/ISH 10-year CVD risk bands.

3.3 Results and discussion

3.3.1 Risk-factor prevalence in rural India

The mean distance of each village from Bhimaravam was 26.91±9.94 km. The average population per village was 3936±1636 people, out of whom 1377±573 were eligible for the baseline study. Overall, 74,402 inhabitants from the 54 villages were eligible to participate in the study. Out of these 74,402 inhabitants, 10,332 (or 13.9%) could not be screened and 1816 (or 2.4%) did not consent to participate. Data collection was performed from 62,254 participants overall, equivalent to 1152±460 participants on average from each village. The mean of the percentage of participants from whom data was collected out of the total eligible and consenting participants in each village (or response rate) was 84.18±2.83%. The diabetes status could not be ascertained owing to missing/corrupted data from 60 participants, and these patients were excluded during data analysis.

The final dataset comprised *62194 participants from 54 villages*. Table 3.2 summarises the data collected using the *SMARThealth* CDSS tool. Our target population is middle aged, with the majority (16320 or 56.0%) of men engaged in manual labour (aquaculture and agriculture are major drivers of the economy in this region [79]) and the majority of women (18441 or 55.7%) being home makers. Only a small fraction of the population (4.3%) was educated at class 11 or beyond (equivalent to A-levels).

Table 3.2 Population characteristics from the WG district, Andhra Pradesh (N=62194). Statistical significance testing was performed to investigate gender-stratified differences for key risk factors using the Wilcoxon signed-rank test. All risk factors showed statistically significant difference except family history of myocardial infarction (MI) and suspected diabetes.

Feature	**Male**	**Female**	P value
Number of subjects % (n)	46.78 (29097)	53.22 (33097)	<0.01
Age, mean±std, (years)	54.9±11.22	53.4±10.82	<0.01
Current Smoker,%(n),	41 (11929)	5.1 (1693)	<0.01
Treated for hypertension,%(n),	16.13 (4692)	22.78 (7538)	<0.01
Self-reported Hypertension,%(n),	19.3 (5607)	26.5 (8755)	<0.01
Family history-MI,%(n),	6.34 (1846)	6.6 (2183)	0.203
Family history-Stroke,%(n),	8.3 (2420)	7.4 (2441)	<0.01
SBP, mean±std, (mmHg)	123.9±20.82	127.75±23.22	<0.01
DBP, mean±std, (mmHg)	79.05±12.18	79.98±11.22	<0.01
Suspected Diabetes,%(n),	17.82 (5202)	18.42 (6104)	0.093

The level of smoking in males was eight times greater than in females. Similar differences in smoking levels have been observed in previous studies. Jonas et al. [80] studied 4711 subjects aged 30 years and above in rural central India and found that 40.3% of males and 0.02% females were current or former smokers. Applying the criterion of hypertension as defined in Section 3.2.3 indicates that 40.5% of the total population is hypertensive. However, the percentage of participants who reported having treatment for hypertension was a mere 19.6% indicating the disparity between the numbers of those with the condition and those under treatment.

A study in rural Andhra Pradesh in 2009 found that over 80% of people with established CVD did not have access to the cheapest treatments [56]. Challenges for a major public health intervention lie not only in identifying high CVD risk and/or hypertensive participants but also in ensuring treatment is available, besides improving awareness in the target population about cardiovascular disease.

Figure 3.5 illustrates the CVD risk profile of screened participants in the WG district of rural Andhra Pradesh by using the low information WHO/ISH CVD risk charts. If we apply the criteria for high-risk individuals (previously defined in Section 3.2.3), we find that 9864 or 15.8% of the participants have a high risk of CVD. The risk profile is positively skewed, with the majority of participants at low risk (40525 or 65.2%). A fifth of the participants (11,806 or 19%) are in the intermediate risk bands (defined to be 10 to <20% or 20 to <30% with SBP<140 mmHg).

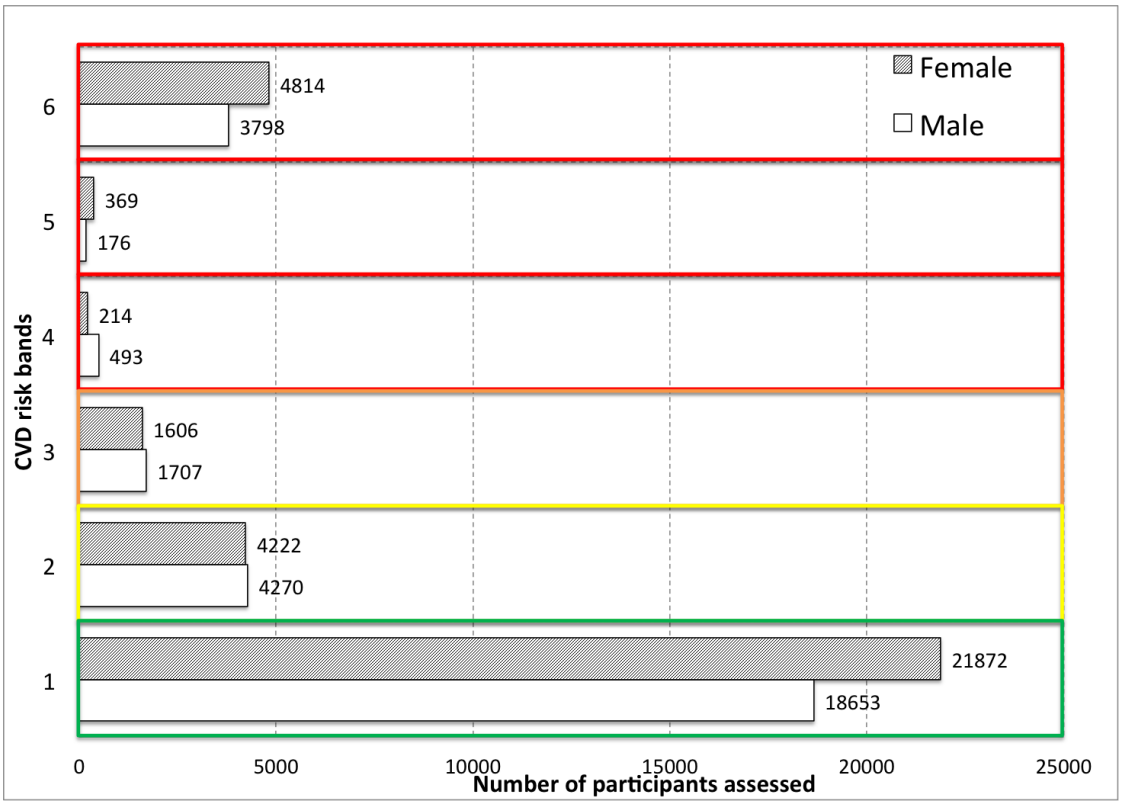
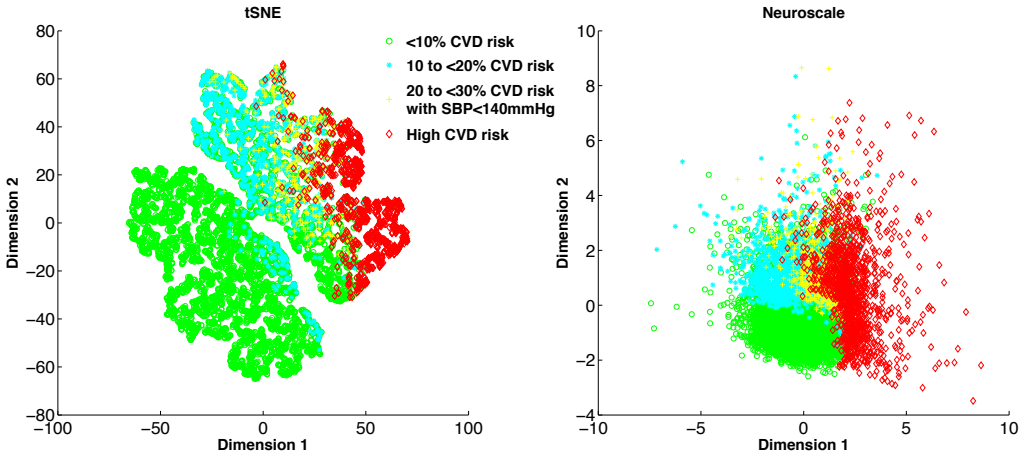


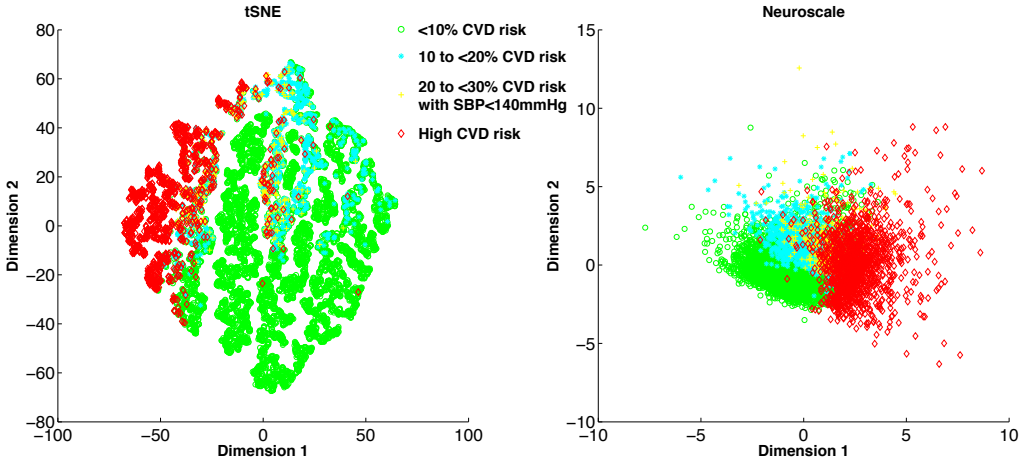
Figure 3.5 Gender-stratified risk profile of screened participants, calculated using the low information WHO/ISH CVD risk charts. CVD risk was stratified as 6 bands in accordance with WHO/ISH risk prediction charts: (1) less than 10%, (2) 10 to <20%, (3) 20 to <30%, (4) 30 to <40%, (5) over 40%, and (6) established CVD and/or clinically high risk.

The two dimensional (2D) representation of the first 10,000 participants (8 villages) free of baseline CVD from Andhra Pradesh are shown via the following plots: Figure 3.6a for male participants and Figure 3.6b for female participants; Figure 3.7a for participants who smoked and Figure 3.7b for those who did not smoke; and Figure 3.8a for participants being treated for hypertension and Figure 3.8b for those not treated. The 2D representation in all plots is labelled by the CVD risk band according to the LI WHO/ISH CVD risk prediction charts.

On comparison of the two-dimensional representations of Neuroscale and tSNE, it may be observed that both are effective in largely separating high CVD risk participants from



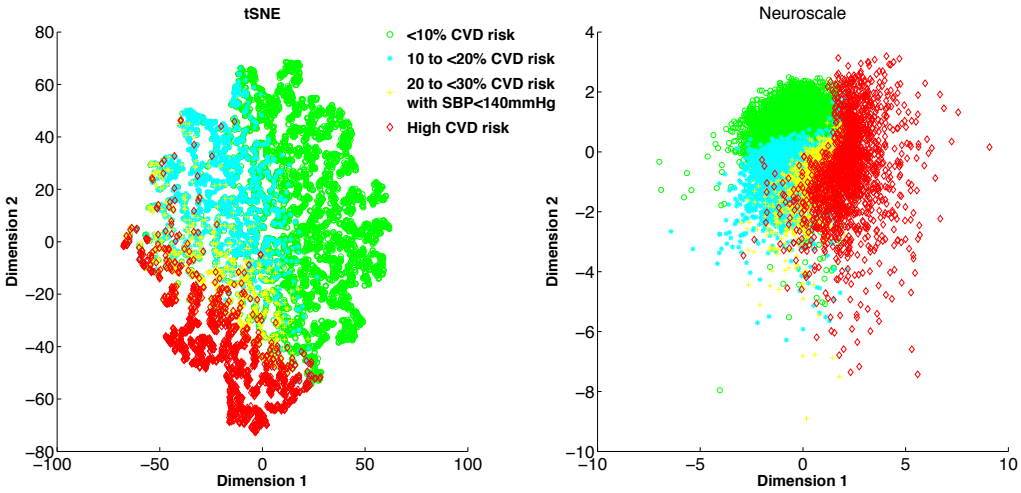
(a) Plot representing male participants.



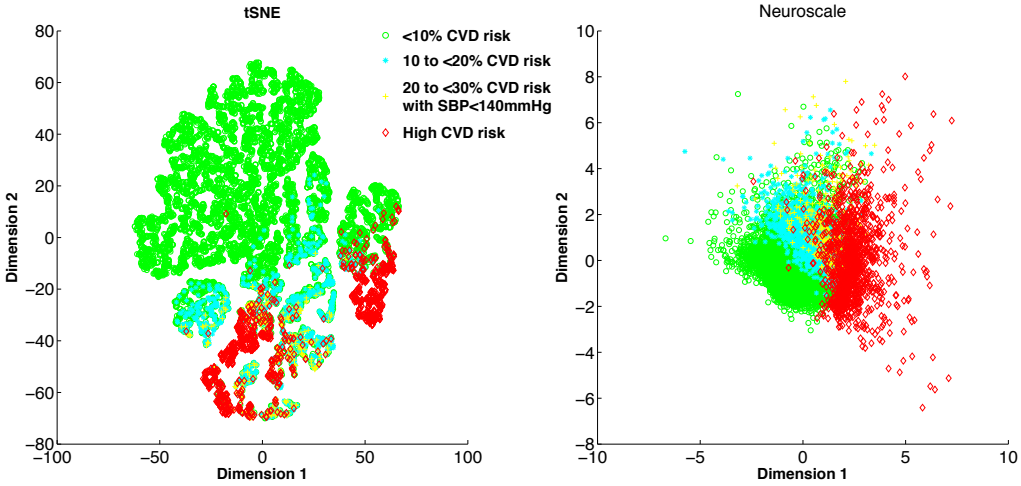
(b) Plot representing female participants.

Figure 3.6 2D representation when data is split by gender.

low CVD risk participants. Neuroscale tends to produce a homogeneous cloud when split by any of the 3 variables (Figures 3.6, 3.7, 3.8). tSNE, on the other hand, tends to group points to form 'islands'. Let us consider the example of male smokers (Figure 3.6a) where both techniques place high CVD risk participants (coloured in red) to be on the right hand side of the plot and low CVD risk participants (coloured in green) on the left hand side. However, tSNE generates a margin of separation between the two risk bands. This tendency to group and to produce 'islands' of points has been consistently observed in other problems where tSNE has been applied [76]. From the visualisation, it is clear that the separation



(a) Participants who were smokers.

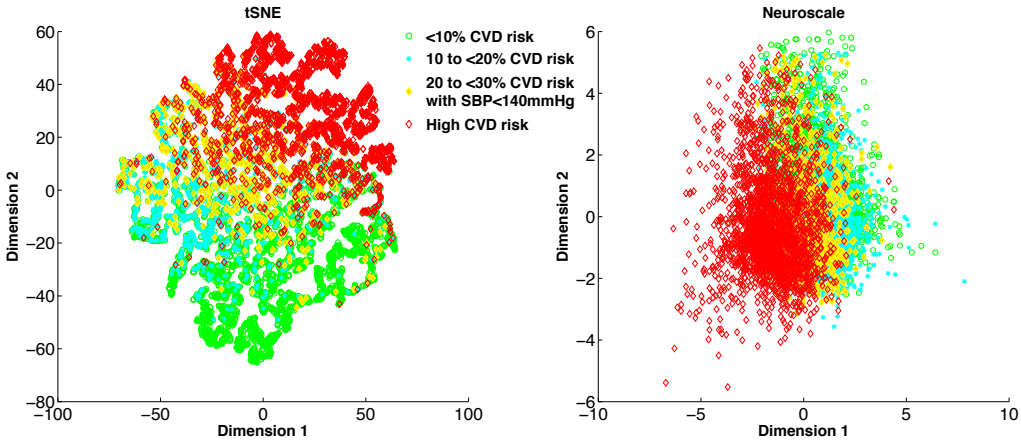


(b) Participants who were non-smokers.

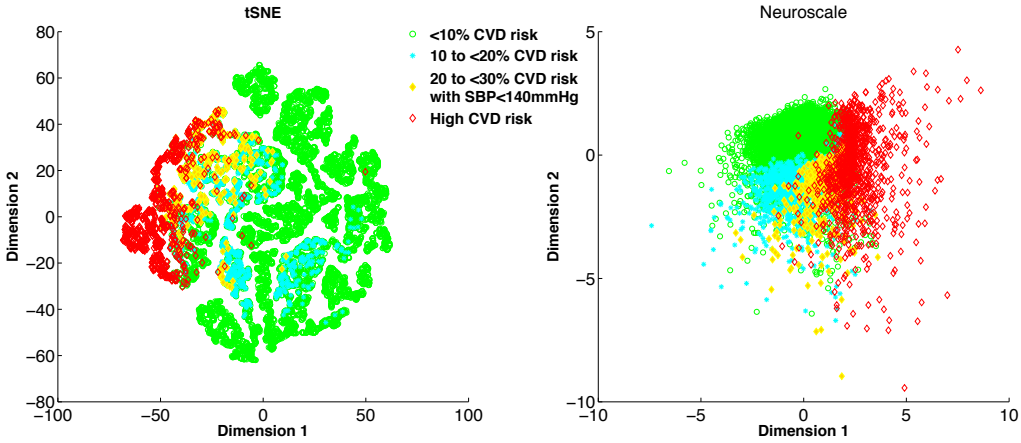
Figure 3.7 2D representation when data is partitioned by those participants who were smokers and those that were non-smokers.

between high CVD risk and low CVD risk participants is distinct. This prompts the question of whether unsupervised exploration can be used for identifying high CVD risk participants, and this will be revisited in Chapter 7.

The difference observed in the low dimensional representation is markedly different when split by gender (as exemplified by Figure 3.6) than when split according to treatment for



(a) Participants on treatment for hypertension



(b) Participants not on treatment for hypertension

Figure 3.8 2D representation when data is partitioned by those participants that were treated for hypertension and those that that were not treated.

hypertension (as exemplified by Figure 3.8). This requires more exploration into whether the gender influences the WHO/ISH charts more than other predictors, which will be explored in the next chapter.

3.4 Conclusion

In this chapter, we have acquired a large dataset of CVD risk factors from 54 villages in rural Andhra Pradesh through the two phase data collection process. In particular, we have presented the use of our *SMARThealth* mHealth CDSS to achieve the task with minimal modifications to enable scale-up. The distribution of 10-year CVD risk was positively skewed. Out of 62,194 participants, it was found that 9864 or 15.8% were at high risk of CVD, 11,806 or 19.0% were at intermediate CVD risk, and 40,525 or 65.2% were at low risk of CVD. The prevalence of key CVD risk factors in rural Andhra Pradesh was discussed. In particular, the level of smoking in males was eight times greater than in females. There was also a substantial difference between those participants who reported having treatment for hypertension (12,190 or 19.6%) versus those were computed to be hypertensive by the application of clinical guidelines (25,188 or 40.5%).

Data obtained from Andhra Pradesh was visualised using a more recent visualisation technique, namely the Barnes-hut variant of the t-Stochastic Neighbour Embedding technique (tSNE), and with a standard non-linear technique using neural networks called Neuroscale. Both techniques could segregate high CVD risk participants well in the low-dimensional space, which may suggest further exploration with unsupervised techniques to identify those at high CVD risk. Also, splitting the data in terms of gender produced markedly different plots in the low-dimensional space as compared to the splits by smoking status or treatment for hypertension. Further investigation of the importance of gender to the WHO/ISH CVD risk prediction charts is performed in the next chapter.

Chapter 4

Investigation of risk prediction algorithms for CVD

4.1 Introduction and rationale for CVD risk prediction

The risk prediction algorithm for CVD is at the core of our clinical decision support system. This chapter firstly introduces current CVD risk prediction algorithms in practice, their methodology, and reported performances. The utility of these algorithms is in the identification of those individuals who are substantially more likely to develop or die of CVD. To build a CVD prediction algorithm, we first require data with recorded outcomes in order to understand the disease incidence rates [81]. Secondly, we need adequate *features* or risk biomarkers that are associated with the occurrence of the disease. Thirdly, statistical techniques are required to build a prediction model which must be evaluated for efficacy. Finally, evaluation metrics must justify the need for the model in clinical practice [81]. This chapter presents an outline of different approaches for predicting CVD risk that will be considered in succeeding chapters. These are described within the practical constraints of the necessary criteria mentioned above, such as the lack of recorded outcomes for rural India.

Conventional clinical practice used to determine CVD risk by inspecting risk factors such as blood pressure and total cholesterol individually [82]. However, it was later realised that a multi-factorial, 'absolute' risk approach was more beneficial [83]. Absolute risk is quantified as the probability of developing CVD within a fixed time-frame, usually within 5 or 10 years. For instance if the 10-year risk of an individual is 25%, it means that given this level of risk factors, the individual can be told that 25 out of every 100 people will experience an adverse CVD event in 10 years. The justification for the 10-year period has primarily been based on effectiveness, affordability, and the safety of commencing drug therapy to mitigate the risk [81]. Evidence has showed that on account of risk factor interactions, a moderate reduction in several CVD risk factors is more beneficial than a major reduction in a single risk factor [84]. Also modifiable elements such as a change in lifestyle, diet/nutrition, or physical activity can lead to an overall reduction in CVD risk and/or to a lowering of multiple CVD risk factors. In clinical practice, the probabilities derived through CVD risk prediction enable the quantification of associated risk. Those at high risk should receive more intensive treatment, as indicated by guidelines such as the Fourth Report of the National Cholesterol Education Program's Adult Treatment Panel (ATP-IV) [85]. Other advantages of risk prediction include raising population awareness of CVD, and communicating severity (or non-severity) of a patient's current lifestyle or adherence to therapy [81].

4.2 CVD risk prediction algorithms in the literature

4.2.1 Overview

Whilst there are several algorithms for prediction of coronary heart disease (CHD) in the literature, a limited number of them predict CVD. An overview of the major risk scores that determine 10-year risk is presented initially. Subsequently, the technical background for the

methods commonly used to develop risk models will be discussed. The reported evaluation metric for the risk prediction algorithms is the Area Under the Curve (AUC). The definitions for the confusion matrix are illustrated in Figure **??**.

		Observed outcome		
		Presence of CVD	Absence of CVD	Prevalence = $\frac{\sum Presence\ of\ CVD}{N}$
Expected value	Presence of CVD	TP	FP	PPV = $\frac{TP}{TP+FP}$
	Absence of CVD	FN	TN	NPV = $\frac{TN}{FN+TN}$
		Sensitivity = $\frac{TP}{FN+TP}$	Specificity = $\frac{TN}{FP+TN}$	Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$

Figure 4.1 Confusion matrix explaining the discriminative metric based on the relationship between the expected value and observed outcome. CVD indicates 10-year cardiovascular disease risk. PPV stands for positive predictive value; NPV - Negative Predictive Value

1. **Framingham risk scores (FRS)**

   The Framingham risk scores are based on long-term prospective studies (namely the Framingham original cohort and offpsring cohort) from the general population residing in Framingham, Massachussets, USA. Three risk scores have been developed. The first risk score by Anderson et al. [86], herein is referred to as FRS-1, was based on data from 5573 participants who were followed from baseline examination between 1968 and 1971. A parametric model following a Weibull distribution [1] was fitted to obtain a risk model. The second risk score (FRS-2) by Wilson et al. [87] was developed for coronary heart disease (CHD) using 5345 participants who were followed for 12 years after baseline data collection between 1971 and 1974. Cox regression analysis

[1] The Weibull distribution is a versatile distribution commonly used to model time-to-failure. The scale and shape parameters can be altered to generate distributions where the failure rate is proportional to the power of time.

was performed to obtain a risk prediction model for CHD. The most recent risk score, FRS-3, was developed by D'Agostino et al. [88] using data from 8491 participants who had baseline examinations from 1984 to 1987. Using Cox regression analysis [89], two gender-specific equations were developed - a 'Global' CVD risk score with laboratory predictors and a 'Simple' model using non-laboratory predictors (body-mass index replacing lipids as a covariate). Reported AUCs were 0.76 (men) and 0.79 (women) based on the training set (data on which the model was developed). Bootstrap samples from the training set were used to estimate the degree of over-optimism in model assessment. This was found to be 0.001 for men and 0.003 for women. All risk scores were based on the data from participants aged between 30 and 75 and those with cardiovascular disease or overt CHD were excluded at baseline.

External validation of the performance of the Framingham risk scores from various studies can be found in the article by Siontis et al. [90], who stated that the Framingham risk scores have inferior performance generally in comparison with other CVD risk models. However, the authors also indicate that performance evaluation is susceptible to biases [90]. They go on to suggest that the requirement is for independent studies by authors other than those who developed the original model [90].

2. **QRESEARCH cardiovascular risk (QRISK 1 and QRISK2) scores**

   The QRISK scores were based on the health records of patients who visited their general practice in the UK [2]. QRISK1 used the data from 1.28 million patients while QRISK2 used the data from 531 general practice clinics (2.29 million patients) aged between 35 and 74. The QRISK2 model utilised 355 general practice clinics (of 1.53 million patients) for the derivation dataset and 176 general practice clinics for the validation dataset. The assignment of general practices to either of these groups was

performed randomly. QRISK2 reported AUCs of 0.79 (men) and 0.82 (women) on the validation dataset. This is the recommended CVD risk screening algorithm in the UK by the National Institute Clinical Excellence (NICE) guidelines on lipid modification [91].

3. **Systematic coronary risk evaluation (SCORE)**

   The Systematic Coronary Risk Evaluation (SCORE) algorithm was developed using data from 12 prospective studies across 11 European countries [92]. Sampling was performed at random from the general population and included 117098 men and 88080 women between 40-65 years. A prediction model to determine the risk of CVD mortality was developed using a Cox model with the hazard function following a Weibull distribution. The risk score is presented as colour-coded risk charts, and reported AUCs on the training data are 0.80 for high-risk patients and 0.75 for low risk patients. The SCORE algorithm is recommended in the European guidelines on CVD prevention [93].

4. **ASSIGN-SCORE**

   The Assessing Cardiovascular risk according to the Scottish Intercollegiate Guidelines Network to assign preventative treatment (ASSIGN-SCORE) algorithm was developed using the data from 6540 men and 6757 women in Scotland aged between 30 and 74 [94]. Cox regression analysis was performed to develop a prediction model for the risk of CVD events. Reported AUCs are 0.73 (men) and 0.77 (women). The method of internal validation used is unclear [95].

5. **Prospective Cardiovascular Munster (PROCAM) scores**

   The PROCAM risk score is based on a prospective study in Germany from 1978 to 1995 [96]. It included data from 18460 men and 8515 women aged between 20 and 75 and calculates the risk of major coronary events, and the risk of cerebral ischemic

events individually. Besides a Cox-Weibull approach, exploratory analysis with neural networks was also performed. The reported AUCs are 0.82 for coronary events and 0.78 for cerebral ischaemic events. It is not clear if any validation set was used for the reported AUCs [95].

6. **WHO/ISH risk prediction charts**

   The World Health Organization/International Society for Hypertension (WHO/ISH) CVD risk prediction charts are prescribed for 14 different epidemiological sub-regions in the world [37]. Two versions of the WHO/ISH charts exist - a low information version (LI-WHO) and a high information version (HI-WHO). Limited information exists on the methodology used to develop these charts [97]. However, the risk prediction charts are recommended by the WHO guidelines on CVD prevention [38], which have been adopted by different countries due to lack of population-specific models.

7. **Reynold's risk score**

   The Reynolds risk score focuses on CVD risk models for women using data from the Women's Health Study, which was a nationwide cohort of predominantly women from USA who were free of cardiovascular disease and cancer at commencement of the study in 1992 [98]. The Reynolds risk score utilises the data of 24,558 women and from 10,724 men aged between 40 and 80. This algorithm was specifically designed for women. The data was split into a derivation set (16400 women) and validation set (8158 women). It predicts incident myocardial infarction, stroke, coronary revascularization, or CVD death [97]. Reported AUCs are 0.70 (men) and 0.80 (women) on the validation set.

A summary of the covariates used for various CVD risk prediction models is given in Table 4.1. We observe that age, gender, and systolic blood pressure have been utilised in all

the prediction models while diabetes and current smoking status are widely used also.

Table 4.1 Covariates of existing CVD risk predictions techniques in the literature. Most commonly used predictors are age, gender, systolic blood pressure, diabetes, and current smoking status. Abbreviations used are described in the footnote[2]

Risk score/ Parameters	FRS1	FRS2	FRS3	HI-WHO	LI-WHO	SCORE	ASSIGN-SCORE	QRISK1	QRISK2	PROCAM	Reynolds score
Age	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gender	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SBP	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DBP	✓	✓									
Hypertension		✓									
Diabetes	✓	✓	✓	✓	✓		✓	✓	✓	✓	
Current smoking status	✓	✓	✓	✓	✓	✓		✓	✓	✓	
#cigarettes							✓				
TC			✓	✓		✓	✓				✓
HDL			✓				✓			✓	✓
LDL		✓								✓	
Elevated cholesterol		✓									
Anti-hypertensive treatment		✓	✓					✓	✓		
FH-CHD							✓	✓	✓		
FH-premature MI											✓
TC/HDL ratio	✓					✓		✓	✓		
ECG-LVH	✓										
$I_{Age-gender}$	✓										
$I_{Diabetes-gender}$	✓										
$I_{ECGLVH-gender}$	✓										
Deprivation index							✓	✓	✓		
BMI								✓	✓		
Ethnicity								✓			
Rheumatoid arthritis									✓		
CKD									✓		
AF									✓		
HbA1c											✓
hsCRP											✓

Siontis et. al (2012) carried out a systematic review of the above mentioned models. They studied 20 articles in the literature and made 56 pairwise comparisons [90]. The authors found that only 10 of the 56 comparisons exceeded a 5% relative difference according to the AUC metric but observed inconsistency in the relative prognostic ability of the popular

[2]Systolic blood pressure (SBP); Diastolic blood pressure (DBP); High density lipoprotein (HDL); Low density lipoprotien (LDL); Family history of CHD (FH-CHD); Myocardial Infarction (MI); Electrocardiogram- Left Ventricular Hypertrophy (ECG-LVH); Chronic kidney disease (CKD); Atrial fibrillation (AF); Haemoglobin A1c (HbA1c); High sensitive C-reactive protein test (hsCRP); and *I* stands for interactions between features

risk models. They found that the results were often biased due to optimism and outcome selection and recommended more evaluations to be done by investigators other than the authors of the models. The Framingham algorithm remains popular internationally and has been adapted for use by many countries for their specific populations [97]. The next section presents a detailed insight into the Framingham study and relevant technical background on the Framingham risk prediction models.

4.2.2 The Framingham heart study

With the advent of World War II, CVD established itself as a major cause of mortality, with 1 in 3 American men developing CVD before reaching 60 years of age [99]. As a result, epidemiological investigations became necessary to understand the aetiology of CVD on which little was known at that time. This found support with a number of people who preferred a primary preventative approach as laboratory and clinical research was not able to yield adequate information at that time [100]. The Framingham Heart Study (FHS) [99] was born, and became a landmark, long-term cohort study to determine the aetiology of CVD. It began in 1948 and since then, has enrolled three generations of participants from the town of Framingham, Massachusetts. Several articles have detailed the background and design of the FHS including its scope and limitations [99]. A brief summary of the study to date is given in Table 4.2.

Technical background

In the literature, frequently used techniques for CVD risk prediction are based on survival analysis [97]. This encompasses models that are either non-parametric, semi-parametric, or fully parametric. Let T be a random variable representing the time to a CVD event. This includes outcomes related to myocardial infarction (MI), Coronary Heart Disease (CHD), CHD death, stroke (including transient ischaemic attack), peripheral vascular disease, and

Table 4.2 Table summarising the progression of the Framingham Heart Study

Cohort	Duration	Description	Number of participants	Established prediction models
Original	1948 - present	Random sample of 2/3 of the adult population in Framingham	5209 (2336 men; 2873 women)	Anderson et al. (1991)
Offspring	1971 - present	Established to study young adults, who were offsprings of the Original cohort	5214 (2483 men; 2641 women)	Wilson et al. (1998) D'Agostino et al. (2008)
III generation	2001 - present	Participants had at least one parent in the Offspring cohort	4085 (1913 men; 2182 women)	

death from CVD. The distribution of T may be described by its probability density function $f(t)$, and cumulative distribution function $F(t)$.

$$F(t) = \int_0^t f(u)du \tag{4.1}$$

The survival function, which describes the probability that an individual will survive beyond time t, is complementary to the cumulative distribution function.

$$S(t) = P\{T \geq t\} = \int_t^\infty f(u)du \tag{4.2}$$

If $F(t)$ is absolutely continuous, then $S(t)$ is equal to 1-$F(t)$. The hazard function captures the instantaneous survival rate, or the probability that an individual will survive in the next

instant given the person survives to t.

$$\begin{aligned} h(t) &= \lim_{dt\to 0}\Big[P\frac{(t \le T < t+dt \mid T \ge t)}{dt}\Big] \\ &= \lim_{dt\to 0}\frac{1}{dt}\Big[\frac{P(t \le T < t+dt)}{P(T \ge t)}\Big] \\ &= \lim_{dt\to 0}\Big[\frac{F(t+dt)-F(t)}{dtS(t)}\Big] \\ &= \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \end{aligned} \tag{4.3}$$

The cumulative hazard function $H(t)$ can be in simple terms thought of as the accumulation of hazard over time.

$$\begin{aligned} H(t) &= \int_0^t h(u)du \\ &= \int_0^t -\frac{S'(u)}{S(u)}du \\ &= -logS(t) \end{aligned} \tag{4.4}$$

Hence,

$$\begin{aligned} h(t) &= -\frac{d}{dt}log(S(t)) \\ S(t) &= exp(-H(t)) \end{aligned} \tag{4.5}$$

Non-parametric models The objective in survival analysis is to handle censoring (a situation in which an observation is only partially known) and evaluate the survival function $S(t)$. The simplest method is a non-parametric estimate of $S(t)$ by using precise event and censoring times, and is called the Kaplan-Meier estimate [101]. To compute the survival function, the risk sets of all subjects being studied are determined at the time of an event. At times with no events, the survival probability is considered to be 1. If n_t individuals are being studied at time t with d_t events occurring at that instant, then the estimated risk r_t and

survival function can be computed as follows:

$$r_t = \frac{d_t}{n_t}$$
$$S(t) = \prod_{t_i \leq t} (1 - r_{t_i})$$

where t_i represents the times at which a CVD event occurs.

The cumulative hazard function can be estimated through the Nelson-Aalen estimator [102] given by:

$$r_t = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{4.6}$$

Semi-parametric models The most common methodology to predict failure in survival analyses is through a Cox regression [103]. The main assumption in the Cox model is one of proportional hazards, implying that the *ratio* of hazards between the two groups (exposed and unexposed to CVD events) is constant over time even though the underlying hazards may change. The proportional hazards assumption can be tested in different ways. A first line approach is to inspect whether the survival curves cross each other. Other tests for assumption can be found in the article by Persson (2002) [104].

If $\mathbf{X} = \{x_1, x_2 \ldots x_k\}$ represent CVD risk factors (such as age, SBP) for an individual and $\beta = \{\beta_0, \beta_1 \ldots \beta_k\}$ denote the coefficients, a linear function of the risk factors can be derived as:

$$h_j(t) = h_o(t) \exp(\Omega_j) \tag{4.7}$$

where

$$\Omega = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{4.8}$$

and $h_o(t)$ is the baseline hazard with j representing a specific patient. If we assume a single covariate with a value of 1 in the exposed group and 0 in the unexposed group, then the hazard ratio of the two groups is proportional to $exp(\beta)$ and does not depend on t.

$$h_1(t) = h_o(t)\exp(\beta.1) = h_o(t)\exp(\beta) \tag{4.9}$$

Hence, the Cox model is a linear model of the log of the hazard ratio [105]. The cumulative hazard may be estimated using the Nelson-Aalen estimator [102]. The fact that no assumption about the baseline hazard following any particular distribution makes the Cox model semi-parametric and advantageous.

The most recent risk scores from the Framingham study (FRS-2 and FRS-3) used age-adjusted, gender-specific Cox proportional hazards models. FRS-3 by D'Agostino et al. [88] employed a Cox regression model on data from the Original cohort and Offspring cohorts. Two versions of the algorithm were presented - a simplified or low information model, and the high information or main model. The following predictors were used in the high information model:

- Age
- Diabetes
- Treated or untreated Systolic Blood Pressure
- Total Cholesterol (TC)
- HDL Cholesterol (HDL)
- Body-Mass Index (BMI)

The simplified model employed BMI instead of lipid measurements (TC and HDL).

A 10-year CVD risk was computed as:

$$P(CVD_{men}) = 1 - 0.94833exp(\sum \beta \mathbf{X} - 26.0145) \tag{4.10}$$

$$P(CVD_{women}) = 1 - 0.88431exp(\sum \beta \mathbf{X} - 23.9388) \tag{4.11}$$

Parametric models In a fully parametric model, $S(t)$ is assumed to follow one of a choice of mathematical functions. For instance, if the hazard is constant over time, then $S(t)$ is fitted to an exponential distribution. Other choices of distributions include gamma, log-normal, log-logistic, and Weibull [106]. The proportional hazards assumption is satisfied by the Weibull and exponential distributions. All the above distributions, however, satisfy the accelerated life assumption, the premise of which states that all patients have the same shape of survival curve, and move faster or slower on the curve according to their co-variates.

Anderson et al. [86] used a parametric approach by modelling the survival function to follow a Weibull distribution. 5573 participants aged between 30 and 74 from both the Original and Offspring cohorts were chosen. They were free of CVD and cancer at the time of recruitment into the study. The authors included two other parameters (θ_0 and θ_1) as a linear function of Ω. Let us define parameters ρ and u as

$$log_{10}\,\rho = \theta_0 + \theta_1 \Omega \tag{4.12}$$

$$u = \frac{(log_{10}(t) - \Omega)}{\rho} \tag{4.13}$$

For a given value of Ω and ρ, we can calculate the survival function to be:

$$P(T > t) = P\left\{\frac{log_{10}(T) - \Omega}{\rho} > u\right\} = e^{-e^{u}} \tag{4.14}$$

The survival function follows a Weibull distribution and parameters are estimated using maximum likelihood. The authors performed selection of covariates in two steps: covariates that model age well were chosen for each gender (e.g. log_{10}*(age).female*) and subsequently additional risk factors were added.

4.2.3 Diabetes-specific CVD risk scores

Diabetes-specific risk models are made to overcome one of the main disadvantages of the models described above - the exclusion of HbA_{1c} and diabetes duration as continuous risk factor variables. It has been reported that the risk models for type 2 diabetes, are specific in nature as these patients experience a risk of CVD two to four times higher than the non-diabetic population, as studied by Gu et al. [107], who performed their analysis with data samples only from the USA .

The United Kingdom Prospective Diabetes Study (UKPDS) developed another risk engine that includes HbA_{1c} and diabetes duration, as well as other risk factors such as systolic blood pressure, smoking and cholesterol levels. This diabetes-specific risk model gives a 10-year risk estimate of developing myocardial infarction, stroke and CVD for patients with type 2 diabetes [108]. Coleman et al. (2007) evaluated the Framingham, SCORE and DECODE risk equations against the UKPDS data for patients with type 2 diabetes. They concluded that the risk equations provide an unreliable risk estimate for fatal CVD and coronary heart disease in the presence of type 2 diabetes, and furthermore emphasised the requirement for a validated diabetes-specific risk calculator [109].

The second version of the UKPDS risk engine estimates the risk of CHD and stroke separately. In the third version however, equations have been derived that estimate CVD risk directly [110]. Simmons et al. [110] examined the performance of this version of the UKPDS risk engine and of the Framingham risk equations to compute CVD risk estimates in the following cohort:

- Individuals who are already diagnosed with diabetes.
- Individuals with non-diabetic hyperglycemia (Hb$A_{1c} > 6.0\%$)
- Non-diabetic individuals with normoglycemia (Hb$A_{1c} < 6.0\%$)

It was found that both risk equations overestimated the CVD risk and there was no significant difference in their ability to discriminate between different risks of CVD events.

4.2.4 Ethnicity and variability in population for CVD risk assessment

In the assessment of CVD risk, it is important to consider the patient's ethnic group because of differences in the incidence and prevalence of CVD in different ethnic groups [111], as well as to improve the risk assessment outcome. Hippisley-Cox et al. [2] developed a CVD risk prediction algorithm taking into account ethnicity and social deprivation factors. They used the QRISK2 algorithm, which was built on the QRISK1 algorithm and included more risk factors such as self-assigned ethnicity, absence of type 2 diabetes and body mass index. Table 4.3 provides examples comparing the QRISK2 and the Framingham equations.

The authors of QRISK2 reported that their algorithm showed good discrimination and calibration upon validation. Some of their key findings were:

- The risk of CVD was higher than previously thought for the South Asian population, as compared to Caucasian patients.
- The results showed evidence that merging people of South Asian origin into one category was deceptive. For instance, one of the results obtained compared the level of

Table 4.3 Comparison of classification results of the QRISK2 and NICE modified Framingham equations presented by Hippisely-Cox et al. [2]. The NICE guidelines do not include Type 2 diabetes, which is therefore highlighted for only illustrative purposes.

Age (years)	Ethnic group	Family history	Systolic blood pressure	BMI	Cholesterol/HDL ratio	Smoker	Treated hypertension	Type 2 diabetes*	Chronic kidney disease	Townsend score†	Framingham score 10 year risk (%)	QRISK2 10 year risk (%) (95% CI)
Men												
65	Indian	Yes	100	24.7	3.3	No	No	No	No	5	17	31.3 (30.9 to 31.7)
54	Bangladeshi	No	142	27.0	4.2	No	Yes	No	No	10	17	23.5 (22.8 to 24.1)
54	Black African	No	150	21.0	7.3	No	No	No	No	4	23	9.0 (7.7 to 10.3)
55	Indian	No	156	27.0	4.7	No	No	No	No	−4	24	12.7 (12.2 to 13.2)
65	Caribbean	No	146	29.1	5.4	No	No	No	No	4	26	14.8 (14.2 to 15.5)
42	White	Yes	132	36.0	5.3	Yes	Yes	No	No	11	17	35.2 (34.9 to 35.5)
Women												
64	Indian	No	130	23.1	5.3	No	Yes	No	No	5	12	24.7 (24.4 to 25.0)
60	Bangladeshi	No	132	36.0	4.3	No	Yes	No	No	11	9	21.1 (20.6 to 21.6)
48	Pakistani	Yes	140	33.2	4.5	No	Yes	No	No	8	9	26.1 (25.7 to 26.4)
58	White	No	154	34.0	3.4	Yes	Yes	No	No	10	16	21.4 (21.3 to 21.5)

BMI=body mass index; HDL=high density lipoprotein cholesterol.
*NICE lipid modification guideline does not include diabetes so this is for illustrative purposes only.
†Interval score ranges between −6 (most affluent) and 11 (most deprived).

adjusted risk for South Asian patients with Caucasian patients as the reference. It was found that the adjusted risk, which took into account many factors including social deprivation and the presence or absence of diabetes, was 45% higher (29% to 63%) amongst Indian men.

- South Asian women were likely to be disadvantaged if existing algorithms were applied for the purpose of treatment. For instance, the Framingham risk equation underestimated their risk of CVD.

- The QRISK2 algorithm considers body mass index and hypertension treatment which are not included in the Framingham score.

- QRISK2 was better than many other CVD risk algorithms in quantifying the risk of CVD in patients with type 2 diabetes. The authors substantiated this with the fact that their study was based on a large nationally representative primary care cohort that included a high proportion of incident events. The risk algorithm also took into consideration vital risk factors such as ethnic group and deprivation [2].

- Collins et al.(2009) compared QRISK1 with the Anderson Framingham model [86] and found that while the former under-predicted CVD risk by 13% for men and 10% for women, the latter over-predicted risk by a greater margin of 32% for men and 10% for women [112]. The authors however did not compare QRISK2 with the NICE modified Framingham equation perhaps due to the timing of publication. However, Hippisley-Cox et al. [2] compared both QRISK2 and the NICE-modified Framingham equations and observed that QRISK2 performed better.

NICE recommended for a long time the Anderson Framingham equation as the first choice for CVD risk assessments in the United Kingdom. However, the NICE Guidance Executive in February 2010 decided against continuing to recommend the Framingham risk equation as the first choice although they stated that it could be one of the possible tools to be used [113]. It is thought that NICE could not find substantial evidence to distinctly recommend one risk equation over the other [114]. There is a lot of debate in the literature about the performance of the different algorithms on different patient groups and the degree of underestimation or overestimation. It is worthwhile considering the variability in population subjects and ethnic minorities because of their inherent propensity for CVD or diabetes. Absence of this could result in health inequalities and imbalance in risk estimates. Many studies in the literature seem to converge in suggesting that both the UKPDS risk engine and the Framingham risk equation are moderately effective at identifying high-risk CVD patients, but are nonetheless equally poor at quantifying the extent of risk.

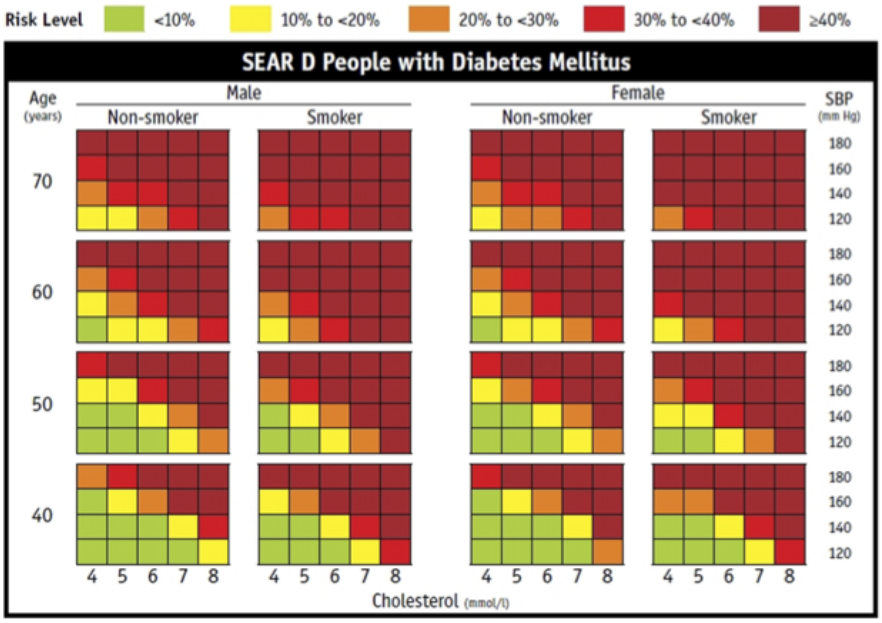
4.2.5 The WHO/ISH risk prediction charts

The WHO/ISH risk prediction charts are a series of colour-coded charts recommended by the WHO guidelines for CVD prevention. Different charts are available for the 14 WHO

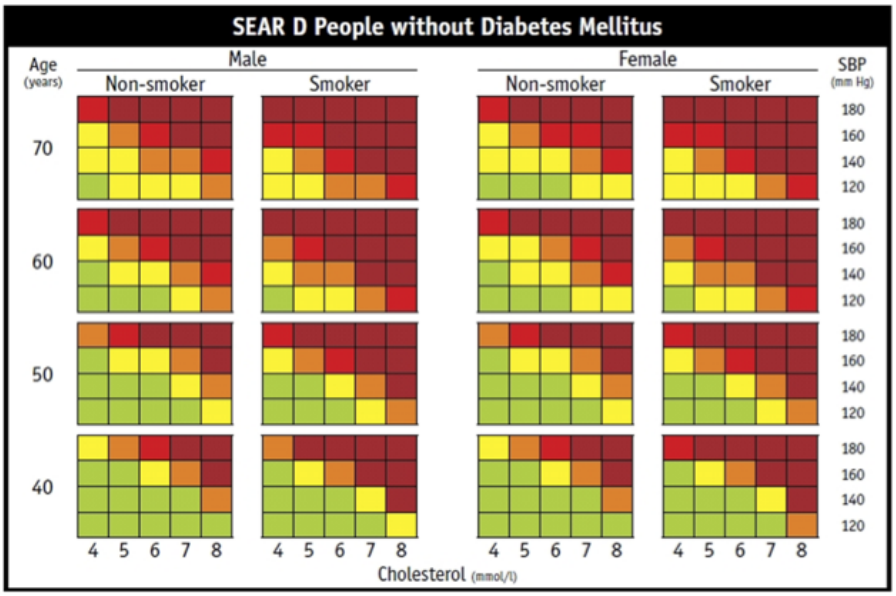
epidemiological subregions around the world [37]. They offer a quantised range for predicting the 10-year risk of fatal or non-fatal cardiovascular event (myocardial infarction or stroke). There are different risk charts based on information on the absence or presence of diabetes. Also, depending on the usage of cholesterol information, a Low Information (LI) or High Information (HI) model of the CVD risk chart is available. The high information charts are used when total cholesterol information is available (Figure 4.2) while low information charts are used where the TC information is not available (Figure 4.3). Common predictors for both models are age, gender, systolic blood pressure, smoking status and presence of diabetes. There are five levels of quantisation in the 10-year risk model; less than 10%; 10 to <20%; 20 to <30%; 30 to <40%; and $\geq$40%.

In the literature there is limited information about the accuracy or validation procedure of the WHO/ISH risk prediction charts. The charts were not developed using prospective or out-of-sample test data and the methods employed differ from other risk estimation functions [97]. Performance metrics such as the classification error between the LI and HI models have also not been reported for the WHO South East Asian Regions D (SEAR-D).

Figure 4.4 presents case studies with the two versions of the WHO/ISH risk prediction charts. The HI WHO charts without Diabetes (Figure 4.2b) highlights the non-linear relationship between gender and other risk factors, which is explained by the examples of Figure 4.4. Non-linear interactions may also exist in the other WHO risk charts; however, it is clear that gender tends to give higher risks for females than males for patients with diabetes as may be observed in both the WHO HI and LI charts (Figures 4.2a and 4.3a respectively). This is also worked out in Case 4 example of Figure 4.4 where the female with diabetes has higher CVD risk than a male with diabetes given that all other risk factors have similar levels. In case of patients without diabetes, females again tend to have higher CVD risk scores for the WHO



(a)



(b)

Figure 4.2 High information (HI) WHO/ISH CVD Risk Charts for the South Asian population. The risk levels are colour coded and as follows: Low risk is indicated by Green (less than 10% risk); Medium risk is shown by Yellow (10% to <20% risk) and Orange (20% to <30% risk); High risk is indicated by Red (30% to <40% risk) and Maroon ($\geq$40% risk).

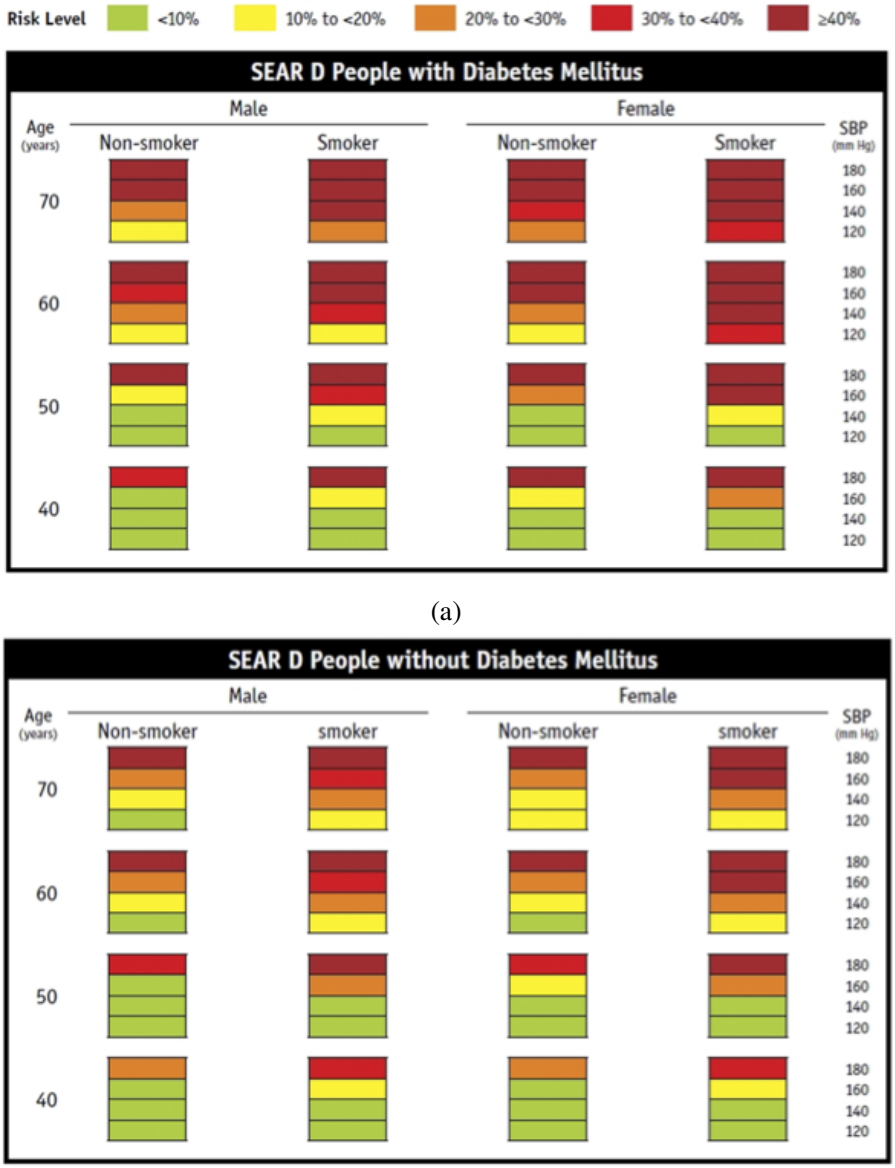


Figure 4.3 Low information (LI) WHO/ISH CVD Risk Charts for the South Asian population. The risk levels are colour coded and as follows: Low risk is indicated by Green (less than 10% risk); Medium risk is shown by Yellow (10% to <20% risk) and Orange (20% to <30% risk); High risk is indicated by Red (30% to <40% risk) and Maroon ($\geq$40% risk).

LI charts (Figure 4.3b) while for the WHO HI charts, males tend to have higher CVD risk scores except for the lone case of 40 year old female smokers with SBP of approximately 180 mmHg and TC levels of 4 mmol/L.

Case 1 Consider the case of a male and female, both aged 45 and active smokers. They are both non-diabetic and have a TC between 8.0 and 8.9 mmol/L with SBP between 120 and 129 mmHg. According to the HI WHO risk chart (Figure 4.2b without Diabetes), the male has a 10-year CVD risk of 10-<20% while the female has a risk of 20-<30%.

Case 2 Consider the case of a male and female, both aged between 70 and 75 and active smokers. They are both non-diabetic and have a TC between 6.0 and 6.9 mmol/L with SBP between 120mmHg and 129 mmHg. According to the HI WHO risk chart (Figure 4.2b without Diabetes), the male has a 10-year CVD risk of 20-<30% while the female has a risk of 10-<20%.

Case 3: Low information version If the risk is computed for the above examples without the cholesterol information, then the 10-year CVD risks are calculated from the LI WHO risk chart (Figure 4.3b without Diabetes). For Case 1, the male and female patients would have a 0-<10% risk. For Case 2, the male and female patients would both have a risk of 10-20%.

Case 4: with Diabetes If the patients considered in Case 1 and Case 2 had diabetes, then the 10-year CVD risks are calculated from the HI WHO risk chart (Figure 4.2a with Diabetes). For Case 1, the male and female patients would have a 30-<40% risk. For Case 2, the male patient would have a risk of 30-<40%, while the female patient would have >40% risk.

Figure 4.4 Examples to illustrate the effect of gender and influence of cholesterol and diabetes in the WHO risk charts.

4.3 Outline of algorithm improvements

We have presented a detailed description of existing risk prediction algorithms and their performance in the previous section. Several risk prediction models have been discussed, yet most of them have been developed specifically for the Caucasian population, and for use in the developed world, thereby not necessarily being applicable to other settings [115]. For instance,the PROCAM score only predicts acute coronary events; QRISK requires inputs like a UK postcode from which deprivation is measured; the Reynolds score was developed for women while the SCORE algorithm was specifically developed for the European population to predict only fatal CVD risk. The WHO/ISH risk prediction charts are the only population-specific algorithm with wide applicability, and are adapted by countries who have not developed their own CVD risk prediction equations using cohort data from their populations. SEAR D, for instance, gives the risk prediction levels for the entire South-East Asian region except Indonesia, Sri Lanka and Thailand. In the succeeding chapters, we will focus on building risk prediction models within constraints of the absence of a gold standard, for our target population in rural India.

4.3.1 Description of datasets

The following datasets will be used in the succeeding 3 chapters:

• *Dataset-1* - N_{d1}=1066 from Andhra Pradesh, India with recorded total cholesterol measurements. This dataset was obtained from the Andhra Pradesh Rural Health Initiative [3] led by the clinical collaborators of the SMARThealth programme, the George Institute for Global Health. This dataset will be used in the next chapter (Chapter 5).

• *Dataset-2* - N_{d2}=3040 from Framingham, Massachusetts, USA that has 10-year follow-up data and outcomes (adverse CVD events) recorded. This dataset was obtained from the Framingham Offspring cohort with permissions from the National Heart, Lung, and Blood

Institute (NHLBI). This dataset will be used in Chapters 6 and 7.

• *Dataset-3* - N_{d3}=62194 from Andhra Pradesh, India. This was collected using our mHealth CDSS, *SMARThealth* and was described at length in Chapter 2. This dataset will be used in Chapter 7.

***Dataset-1* - The Andhra Pradesh Rural Health Initiative**

The Andhra Pradesh Rural Health Initiative (APHRI) is a cross-sectional study of CVD risk factors for 4535 subjects (48.6% male) in rural Andhra Pradesh [3]. Participants from over 20 villages were stratified by age and gender and selected by randomly sampling. Each age and gender stratum could therefore be represented equally (especially the elderly participants who are more vulnerable to a CVD event). Blood samples were drawn from every fourth participant in the study. More details on the APHRI study can be found in the article by Chow et al. [116].

For our analysis, only those subjects who had recorded blood cholesterol and blood glucose measurements were included (N=1066). The characteristics of the selected subjects are summarised in Table 4.4. Statistical significance testing was performed to compare the chosen subjects (N=1066) with the APRHI cohort (N_a=4535) using a two-sample, two-sided t-test for continuous variables, and the Wilcoxon signed-rank test for variables 'smoking' and 'treated for hypertension'. The difference was not statistically significant ($p>0.05$) for all gender-stratified variables in Table 4.4 except male smokers ($p=0.0095$).

***Dataset-2* - The Framingham study**

Data from exam 6 of the Framingham offspring cohort was considered for analysis. The offspring cohort was chosen for analysis as the 3rd generation cohort did not record enough events (65 events out of 3391) while the original cohort dataset was collected in 1958, and in addition to the lack of heterogeneity in the dataset (on account of lower genetic diversity), the availability of parameters like cholesterol was limited during early exam cycles. The choice

Table 4.4 Population characteristics from the chosen subset of APHRI (N_{d1}=1066). Statistical significance testing was performed to investigate this chosen subset from the full APHRI cohort (N_a=4535). Only male smokers were significantly different (p=0.0095 (<0.01); Wilcoxon signed-rank test). BP indicates blood pressure.

Feature	Male	Female
N	48.8% (520)	51.2% (546)
Age, mean±sd, (years)	50.6±14.1	48.2±13.4
current smoking status,%(n)	40.6% (211)	5.1% (28)
Systolic BP, mean±sd, (mmHg)	126.1±20.1	122.4±20.5
Diastolic BP, mean±sd, (mmHg)	77.2±11.5	76.2±10.7
Glucose, mean±sd, (mg/dl)	99.8±28.7	102.4±35.9
Total Cholesterol, mean±sd, (mg/dl)	177.8±40.4	191.1±38.7
Treated for hypertension,%(n)	13.8% (72)	14.8% (81)

of exam cycle was for two major reasons; firstly because exam 6 recorded higher event rates (11.35%) and secondly, because data on family and medical history of the patients were available. Exam 6 consisted of 3434 patients, out of whom 3040 were free of baseline CVD, and 519 variables that included data on pulmonary function tests and respiratory surveys. Key CVD risk factors from exam 6 are summarised in Table 4.5.

Chapter 5 will explore risk prediction through the WHO/ISH charts for SEAR-D in detail. Firstly, the clinical significance of using one version of the chart over the other in the context of CVD will be discussed. Secondly, a point-of-care algorithm to determine the benefit of total cholesterol testing whilst performing CVD risk assessment with the WHO/ISH charts will be presented. Data analysis will be carried out using *Dataset-1*.

Chapter 6 will make use of *Dataset-2* to select highly predictive features from a large list of variables recorded in the Framingham study. The objective is to identify low-cost, highly-predictive features and build prediction models to compare performance against the Framingham risk scores. An approach to a first-order re-calibrated risk model for rural India will be discussed.

Table 4.5 Population characteristics of data from Exam 6 of the Framingham Offspring cohort. Only participants free of baseline CVD were chosen (N_{d2}=3040). BP indicates blood pressure; MI stands for myocardial infarction.

Feature	Male	Female
N	44.67% (1358)	55.33% (1682)
Age, mean±sd, (years)	57.6±9.5	58.0±9.5
Current smoking status,%(n)	7.3% (221)	8.4% (257)
Systolic BP, mean±sd, (mmHg)	129.2±16.6	125.5±18.8
Diastolic BP, mean±sd, (mmHg)	77.0±9.0	72.7±8.7
Diabetes,%(n)	3.4% (102)	3.2% (98)
Total Cholesterol, mean±sd, (mg/dl)	200.7±40.1	211.7±38.3
Family history,%(n) -Diabetes	7.8% (238)	11.5% (349)
Family history,%(n) -MI	5.7% (174)	7.6% (232)
Family history,%(n) -Stroke	4.0% (123)	4.7% (144)

Chapter 7 - The lack of population-specific outcomes leads to difficulties in performing a sophisticated recalibration procedure using techniques such as logistic regression. This motivates the investigation of an unsupervised machine learning approach for identification of high CVD risk patients. *Dataset-2* and *Dataset-3* will be used to demonstrate the validity of a clustering approach to identify high-risk individuals.

4.4 Performance of FRS, WHO/ISH charts, QRISK2 on Indian data

In this section, the performance of three widely used CVD risk prediction algorithms on *Dataset-1* with N_{d1}=1066 patients are studied. The global Framingham CVD risk prediction model by D'Agostino et. al (2008) or FRS-3 [88] comprises a HI version that uses traditional risk factors, and an LI or simplified version that uses non-laboratory based predictors (Equation 4.10). The WHO/ISH risk prediction charts for SEAR-D also consists of a LI and HI

version.

QRISK2 is available as C code. In this analysis, the 2015 update was used (available from http://svn.clinrisk.co.uk/opensource/qrisk2). A MATLAB interface to the C code was created (as technical support for the C code nor coefficients were available). The ethnicity variable was fixed to be Indian. The following variables were unknown and are considered to be absent: atrial fibrillation, Rheumatoid arthritis, stage 4 or 5 chronic kidney disease, and type 1 diabetes. These assumptions were made due to limited availability of data and to conform to the context of the Indian population. QRISK2 includes a variable that accounts for deprivation through a score called the Townsend score. To understand the influence of this variable when examining the performance of the QRISK2 scores on *Dataset-1*, the CVD risk scores were computed at three levels of the Townsend score - maximum deprivation, minimum deprivation, and median deprivation.

Although the QRISK2 scores are used in this analysis to serve as a comparator, it is important to understand its limitations. Firstly, although the QRISK2 algorithm has the advantage of being trained on millions of patients [2], a substantial portion of the training data was imputed (as much as 70%) for certain variables [97]. This has the possibility of introducing bias if the algorithm is to be used in rural India when the imputation methods relied on the statistics of the UK population (e.g. the prevalence of disease).

Secondly, it may be observed that both the FRS and the WHO/ISH risk prediction algorithms, unlike the QRISK2 scores, include a low information version which is better suited to resource-constrained settings. This is because, at point-of-care, besides the cost of data acquisition, a wrong assessment of an important binary/categorical variable (such as stage 4

or 5 chronic kidney disease) can bias the predicted scores considerably.

Thirdly, a variation of the QRISK2 scores across three levels of deprivation is presented. A custom deprivation score, however, needs to be formulated for rural India. This is because the UK score is based on the following four variables obtained from the UK census data: unemployment (lack of material resources and insecurity); overcrowding (material living conditions); lack of owner occupied accommodation (a proxy indicator of wealth); and lack of car ownership (a proxy indicator of income). Our target population in the West Godhavari district in rural Andhra Pradesh, is a relatively well-off community in comparison to southern rural India (which is in turn more well-off than many parts of northern rural India). As described in Chapter 3, the villagers in the West Godhavari district include businessmen (involved in aqua-culture) and landlords (or 'zamindars' who also own vast amounts of farmland), who have considerable wealth, whilst those involved in labour-intensive jobs (such as working on the farm) are impoverished. This thesis will progressively introduce an mHealth platform that will be evaluated through a clinical trial (Chapter 8). The trial data will generate longitudinal information on risk factors that can be compared with risk factors acquired in different locations, and associated with different occupations, which can lead to formulating specific deprivation scores for rural India.

The performance of the HI and LI Framingham models with the HI and LI WHO/ISH risk prediction charts for SEAR-D and the QRISK2 scores was assessed. An important limitation in this exercise is the fact that the WHO/ISH risk prediction charts offer only discrete predictions in 5 risk bands whilst FRS-3 performs continuous prediction. Hence, our risk comparison was stratified into the five categories prescribed by the WHO/ISH charts, and is illustrated in Figure 4.5. This is suboptimal as the FRS-3 authors employed 0% to 6%, 6% to 20%, and >20% as risk categories [88]. However, the aim here was to understand how

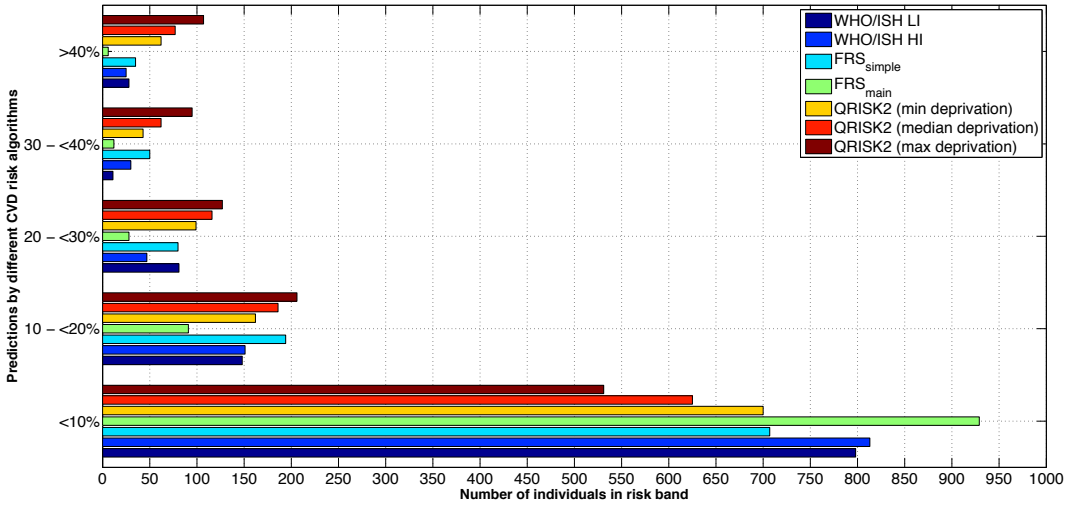


Figure 4.5 Comparison of CVD risk prediction algorithms namely the LI and HI WHO/ISH models, the simple and main FRS-3 algorithm, and the QRISK2 scores on *Dataset-1*. The Framingham main (or HI) model underpredicts risk as compared to the LI (or simplified) model. The HI and LI WHO/ISH charts differ in making prediction between risk bands over 20% and less than 40%. As the deprivation score increases, the number of those classified to be at high CVD risk increases.

the high and low information versions of these algorithms perform on data from our target population.

Figure 4.5 illustrates the differences in risk prediction between a LI (or simplified) and HI (or main) version of the WHO/ISH and Framingham algorithms, as well as across varying deprivation indices of the QRISK2 scores. For the WHO/ISH model, a major difference lies in risk prediction between the risk bands over 20% and less than 40%. It may be observed that as the deprivation level increases for the QRISK2 scores, more people are classified to be at high risk.

The Framingham HI model under-predicts CVD risk as compared to the LI Framingham model with a difference of over 20% of the population (or 222 patients) being in the low risk (0-<10%) range for the former (Figure 4.6). This is subsequently reflected by 103 fewer

patients in the 10-20% risk range, 52 fewer patients in the 20-30% risk range, 38 fewer patients in the 30-<40% risk range and 29 fewer patients in the >40% range for the HI model in relation to the LI model.

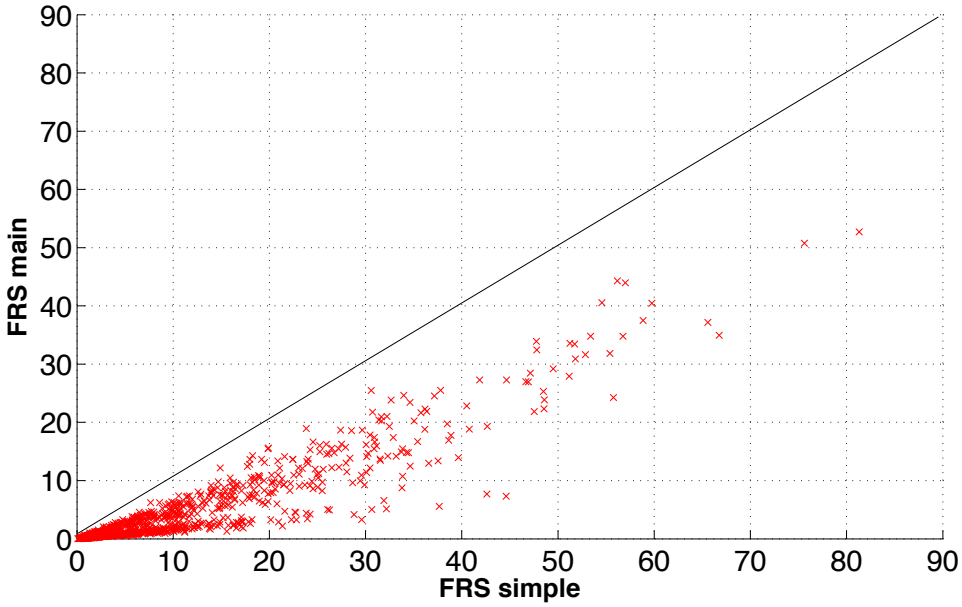


Figure 4.6 Scatter plot of the simple and main FRS-3 algorithm on *Dataset-1*, which is from an Indian population. The diagonal line displayed shows the case of perfect agreement between both versions of the Framingham algorithm. The coefficient of determination (R^2) is 0.86 indicating that 86% of the total variation in FRS-3 main algorithm can be explained by the linear relationship between the FRS-3 simple and FRS-3 main algorithm.

If the Framingham HI and LI models agreed perfectly with each other, they would be along the line connecting the origin to (-90,90) in Figure 4.6. Deviations from this line can be expressed through the coefficient of determination (R^2) which is 0.86 between Framingham HI and LI models. This can be interpreted as follows: 86% of the variation in Framingham HI algorithm can be explained by the linear relationship between Framingham HI and LI algorithms. This indicates that the presence of TC information is important in both the WHO/ISH and Framingham models for classifying CVD risk. These findings need to be

investigated in greater detail in order to understand the clinical relevance.

4.5 Summary

In this chapter, we have introduced the rationale for CVD risk prediction and presented a review of the major CVD risk prediction techniques in the literature. We focused on the widely used Framingham risk scores and the WHO/ISH risk prediction charts that are suitable for rural India. We briefly presented case studies to show the influence of gender and total cholesterol in CVD risk prediction scores calculated using the WHO/ISH risk prediction charts. In particular, the non-linear interactions between gender and diabetes was observed as exemplified by the fact that females tend to have higher CVD risk than males in all cases, except for those patients who have diabetes and who obtained their risk score by specifically using the WHO/ISH high information (HI) chart. Using the APHRI dataset that was also recorded from our target population, we compared the low and high information versions of the WHO/ISH and the Framingham risk prediction models with QRISK2 scores with varying deprivation indices. The Framingham HI model under-predicts risk as compared to the LI model (R^2=0.86).The inclusion of cholesterol information appears to be important when using both the Framingham and WHO/ISH models. Further investigation is required especially with regard to the clinical importance of the mis-classifications for rural India. This will be discussed in detail in the next chapter.

Chapter 5

CVD risk prediction using WHO/ISH charts

5.1 Introduction

We have previously indicated that the prevalence of cardiovascular disease is increasing in the developing world [117]. The Indian subcontinent accounts for the highest rates of CVD globally [118]. Although many algorithms for CVD risk assessment have been developed worldwide, no cohort data is available in these regions for population-specific development of CVD risk models. Currently, the CVD risk prediction charts developed by the World Health Organization and International Society for Hypertension [37] are the only recommended methods for the Indian subcontinent for CVD risk assessment in national guidelines [119]. For instance, in India, interventions for CVD and associated risk factors like diabetes through the National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular diseases and Stroke (NPCDCS) depend on the WHO/ISH South East Asian Regions D charts for CVD risk assessment [1]. SEAR-D countries include Bangladesh, Bhutan, Democratic People's Republic of Korea, India, Maldives, Myanmar, and Nepal [37]. The previous chapter introduced the WHO/ISH risk prediction charts and the case studies that were discussed (see

Sections 4.2.5 and 4.4) revealed that total cholesterol played a key difference in CVD risk predictions using the WHO/ISH charts. We address key issues concerning risk prediction with the WHO/ISH risk prediction charts, specifically for the SEAR-D population and focus on the following:

Objective 1 Quantify the relative error between the LI and HI WHO/ISH risk prediction charts based on data from rural India, specifically highlighting those patients who are likely to be under- or over-treated as a result of choosing one version of the charts over the other.

Objective 2 Use the same dataset to develop a sparse point-of-care algorithm which allows the user to identify patients who are likely to benefit from a TC test when their CVD risk is being assessed through the WHO/ISH risk prediction charts.

We make the assumption that the WHO/ISH HI model is more accurate than the LI model because it uses more information (through the inclusion of total cholesterol, a strong predictor of CVD). There are over 3 million new CVD cases a year in India and so the aim of this analysis is to facilitate more accurate screening of CVD risk. Additionally, we envisage that this study will be of practical relevance to healthcare policy makers and payers, especially in low- and middle-income countries, who may not be able to fund TC tests for the whole population.

5.2 Methods

Dataset-1 from the Andhra Pradesh Rural Health Initiative as described in the previous chapter (Section 4.3.1) was used. A total of 1066 patients had TC and glucose measurements recorded and were considered for analysis.

5.2.1 Risk stratification according to Indian national guidelines

According to the 2009 NPCDCS guidelines, a patient with less than a 10% 10-year risk score should be considered to be at low risk of developing CVD. If the risk score is between 10% to <20%, the patient should be considered to have a moderate risk of developing CVD. Patients with risk scores between 20% to <30% should be considered to be at high-risk while those above 30% should be deemed to be at 'very high risk' of developing CVD [1]. Treatment for high-risk subjects is initiated either if the CVD risk score is above 30%, or if the CVD risk score for an individual is between 20% up to 30% and their systolic blood pressure is also over 140 mmHg [120]. The group of subjects who are at high risk and require treatment is henceforth referred to as T_{HR}. For objective 1, we use the NPCDCS guidelines for the interpretation of the risk range and highlight T_{HR} subjects who are clinically more relevant. For objective 2, we develop a POC algorithm using data from all patients (all categories of risk).

5.2.2 Feature selection

In machine learning, *the curse of dimensionality* is an important consideration. It states that the dimensionality of the feature space is limited by the available data, especially when the number of features is large. A typical example may be genetic data where the data from a handful of patients consists of thousands of features (genes). Consequently, a predictive model with limited training examples tends to be less accurate with increasing features, also known as the Hughes effect [121]. This aligns with the principle of parsimony where we are interested in maximising the predictive power with information from as few features as possible.

A total of 40 features that have been reported in the CVD literature were chosen from the APRHI dataset (listed in Appendix D) for input into a feature selection algorithm, the Maximum Relevance Minimum Redundancy (mRMR) approach [122]. Amongst all features, educational level was the only feature that was ordinal. Since the association of specific levels of education with the outcome was of interest, each category was taken as a feature (which was a binary variable). The mRMR criterion is dependent on the pairwise mutual information between features and class labels. The mutual information (MI) between two continuous random variables X, Y can be defined as

$$MI(x,y) = \int_{x=\infty}^{\infty} \int_{y=\infty}^{\infty} p(x,y).log_b \left(\frac{p(x,y)}{p(x)p(y)} \right) dx\, dy \tag{5.1}$$

where *p(x)*, *p(y)* are the marginal probability densities and *p(x,y)* is the joint probability density function (pdf) of X and Y. The logarithm to the base b is used to denote the unit of MI (for instance, when b=2, the unit is a bit, which is a common way of measuring MI). Equation 5.1 can also be expressed as the Kullback- Leibler (KL) divergence of the joint pdf with the product of the marginals, i.e. $D_{KL}(p(x,y)||p(x)p(y))$. When X and Y are independent, the logarithmic term involved in the computation of KL divergence becomes 0 (as $p(x,y) = p(x)p(y)$), and hence the KL divergence is zero.

The integrals become summations when the random variables A and B are discrete. In practice, the random variables A and B can be a mixture of discrete and continuous variables. In such a case, the KL divergence can be computed by discretising the continuous variable such as through the techniques reviewed in the article by Garcia et al. [123]. Alternatively, the density of a discrete variable can be determined through techniques such as histograms, kernel density estimates, or Parzen windows.

Intuitively, MI can be understood as the amount of information shared between two variables. Unlike correlation which only captures a linear relationship (in the case of Pearson correlation) or monotonic relationship (in the case of Spearman correlation), MI is a better estimate of the dependence between two variables [124]. MI is also commutative, that is *MI(X,Y) = MI(Y,X)*. Hence MI is useful to represent the association between features $\mathbf{f}_{1...m}$ and the response $\mathbf{y}$. This is utilised by mRMR, originally proposed by Peng et al. [122]. mRMR considers the interaction between a feature and the outcome, individually. It penalises the pairwise MI between features (*redundancy*) while maximizing the MI between each feature and the class of outcomes $c \in [0,1]$ (*relevance*). Multivariate interactions are not accounted for by mRMR. However, pairwise feature redundancies are taken into consideration by mRMR.

$$mRMR = \max_{S} \left\{ \frac{1}{|S|} \sum_{\mathbf{f}_i \in S} MI(\mathbf{f}_i; \mathbf{c}) - \frac{1}{|S|^2} \sum_{\mathbf{f}_i, \mathbf{f}_j \in S} MI(\mathbf{f}_i; \mathbf{f}_j) \right\} \tag{5.2}$$

where S is the feature set, MI is the mutual information between feature $\mathbf{f}_i$ and class $\mathbf{c}$ or between features $\mathbf{f}_i$ and $\mathbf{f}_j$. mRMR is an example of a 'filter' method for feature selection, where a statistical criteria forms the basis of ranking features. In this analysis, we use three classifiers for model development (outlined in the next section). In order to perform feature selection that is quick and convenient, as well as independent of using a classifier, the mRMR technique was used. mRMR has been reported to improve model performance through robust feature selection [125] albeit differences in implementation, for instance, in estimation of the joint density, can produce dissimilar results.

To estimate the density of a pdf, a non-parametric method is often used called the kernel desntiy estimate (KDE). If $(x_1, ..., x_n)$ are samples that follow i.i.d (independent and identically distributed) and drawn from a distribution with unknown density f, the KDE may

be estimated as:

$$P_{KDE}(x) = \frac{1}{Nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{5.3}$$

where $K(\cdot)$ is the kernel function (which is non-negative and integrates to 1), and h is the bandwidth of the d dimensional kernel used. If we use a Gaussian kernel (which is common practice), then the kernel function becomes $K(u) = (2\pi)^{-d/2} e^{-\frac{1}{2}u^T u}$. A small bandwidth results in an undersmoothened curve of the estimated density while usage of a large bandwidth results in a oversmoothened curve. The optimal choice of h is dependent on the desired tradeoff between variance and bias. Density estimation in this analysis was performed using Silverman's rule [126], which uses a Gaussian kernel given by

$$P_{KDE}(x) = \frac{1}{Nh^d} \sum_{i=1}^{n} (2\pi)^{-d/2} e^{-\frac{1}{2}(\frac{x-x_i}{h})^2} \tag{5.4}$$

with the bandwidth h in 1 dimension determined by

$$h = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \tag{5.5}$$

where σ is the standard deviation of the samples.

From the top 20 features ranked using this criterion, we selected those features which are clinically relevant features that were either easy to measure, low-cost, or easy to access. The feature set included the following 10 variables: gender, age, current smoker, past history of diabetes, past history of high cholesterol, body mass index, systolic blood pressure, diastolic blood pressure, treatment for hypertension and blood glucose.

5.2.3 Development of model for identifying TC candidates

Data from the 1066 patients in *Dataset-1* were divided into training and testing sets in a ratio of 70:30. The WHO HI and LI risk prediction charts were implemented in the Matlab programming environment. The predicted CVD risk ranges were then obtained from both the HI and LI risk prediction charts applied to the data. We wish to identify the subpopulation that would benefit from a TC test. We perform this by identifying the patients for whom the predictions for LI and HI risk prediction charts differ, and this was the target outcome of our model. Three different classification approaches were applied to the task of identifying these patients from the features identified at the end of the Section 5.2.2 (above), namely Logistic Regression, Support Vector Machine, and Random Forest.

Logistic Regression

Logistic regression (LR) is a probabilistic method for classifying binary outcomes and it is often employed as a standard baseline approach. Let us consider K independent variables $X = \{x_1, x_2 ... x_k\}$ that model the dichotomous dependent variable y. If the probability of a patient i developing CVD is expressed as P_i and the probability of not developing CVD is $P(y_i = 0|x_i) = 1 - P_i$, then

$$log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \sum_{j=1}^{K} \beta_j x_j \tag{5.6}$$

$$\frac{P_i}{1-P_i} = e^{(\beta_0 + \sum_{j=1}^{K} \beta_j x_j)} \tag{5.7}$$

A simple transformation of the above equation leads to

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{K} \beta_j x_j)}} \tag{5.8}$$

If the patient data is assumed to be independent, the coefficients β_j can be estimated through a likelihood function, which can be expressed as

$$L(\beta|\mathbf{x}) = \prod_{i=1}^{N} p_i^{y_i}[1-p_i]^{1-y_i} \tag{5.9}$$

For mathematical convenience, we consider the log of the likelihood function,

$$\mathscr{L}(\beta|\mathbf{x}) = \sum_{i=1}^{N} y_i log(p_i) + (1-y_i) log(1-p_i) \tag{5.10}$$

$$= \sum_{i=1}^{N} y_i(\beta_0 + \sum_{j=1}^{K} \beta_j x_j) + \sum_{i=1}^{N} log(1-p_i) \tag{5.11}$$

The task is therefore to maximise the log-likelihood or minimise the negative log-likelihood for computationally efficiency. The equation can be solved through iterative techniques such as Gradient descent [127] or Newton's method [128] to estimate the β coefficients.

Regularisation Regularisation is performed to reduce overfitting and increase generalisability. It is useful to derive a sparse or parsimonious model that balances model complexity and accuracy, as well as aid the interpretability of the model. This is achieved through the addition of a penalty term which results in shrinking large coefficients. The penalty term (λ) that determines the strength of regularisation, along with a function of the coefficients β, is added to the log-likelihood estimation.

$$\mathscr{L}(\beta|x) = \sum_{i=1}^{N} (y_i log(p_i) + (1-y_i) log(1-p_i)) + \lambda f(\beta) \tag{5.12}$$

$$\tag{5.13}$$

Higher values of the penalty term constrain the β coefficients to be small. Frequently used functional forms of $f(\beta)$ are the L1 norm and the L2 norm. The L1 norm is simply a summation of the absolute values of each coefficient while the L2 norm is the square root of the sum of the squares of each coefficient. Regularisation techniques such as the LASSO, which uses an L1 penalty, when applied, result in a selection of coefficients being zero. According to Hastie, Tibshirani, and Wainwright, "the L1 penalty provides a natural way to encourage or enforce sparsity" [129]. Statistical efficiency of the L1 penalty was discussed by the same authors and they introduced a "bet-on-sparsity-principle". According to this principle, if an underlying true signal was assumed to be sparse, an L1 penalty could be imposed to recover it failing which other methods would not work as well, relative to the Bayes error. Further detail can be found in Chapter 11 of the Statistical Learning with Sparsity book [129].

On performing regularisation, a large λ leads to heavy penalisation of the coefficients, which means a large proportion of the coefficients will shrink to zero. This leads to high bias and therefore underfitting. On the other hand, a small value of λ leads to fewer coefficients being zero and increases the variance, thereby indicating the propensity for overfitting. Regularisation helps to achieve the optimal model selection if the appropriate magnitude for λ can be found. Figure 5.1 explains the bias-variance tradeoff. With increase in model complexity (small penalty), the tendency to overfit increases (high variance, low bias) while an extremely high penalty can lead to underfitting (low variance, high bias).

Bishop [131] provides a detailed explanation of the L1 and L2 norm. In general, given two correlated features, the L1 norm selects one feature from the two while the L2 norm keeps both and shrinks both coefficients. Hence the L1 norm has been known to produce sparse models (i.e. with more zero coefficients) while the L2 norm minimises the prediction error more (because of the quadratic penalisation term) [130] [132]. Here we use the $L1$

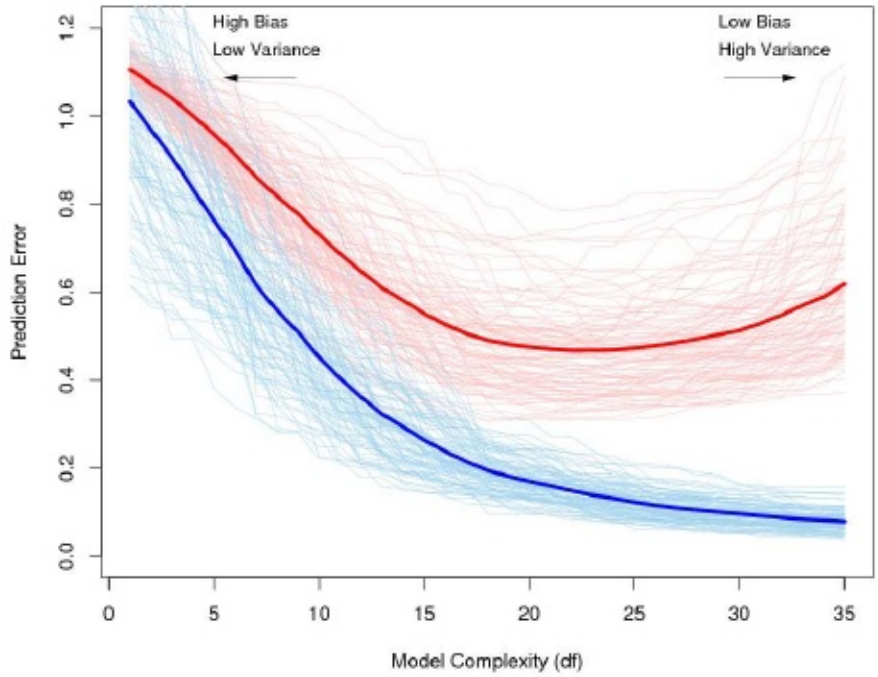


Figure 5.1 Figure illustrating the bias-variance tradeoff (adapted from the Elements of Statistical Learning book by Friedman, Tibshirani, and Hastie [130]).

norm to derive a sparse model because we only have the data for 1066 patients.

Support Vector Machines

Support Vector Machines (SVM) are a class of margin-based classifiers, a concept rooted in statistical theory that links generalisation performance (given unseen test data) and an algorithm's learning capacity (to classify an arbitrary number of points) [131]. A margin can be defined as the smallest distance between the decision boundary and any of the samples [131]. In the case of SVMs, the decision boundary is chosen such that the margin is maximised. Let us assume we are given samples from the positive x_+ and negative x_- classes which are linearly separable in the feature space, as illustrated in Figure 5.2.

If $\mathbf{w}$ represents a line constrained by being perpendicular to the decision boundaries, and we are given an unknown sample $\mathbf{u}$, the projection of $\mathbf{u}$ onto line $\mathbf{w}$ is given by the dot

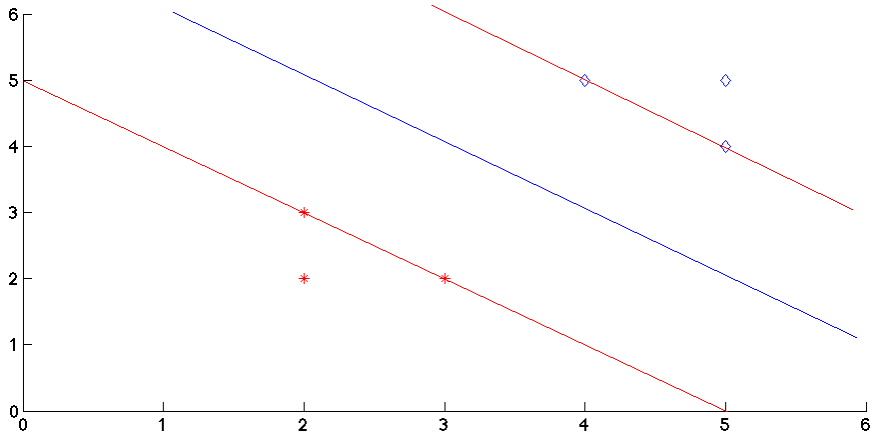


Figure 5.2 Figure illustrating a linearly separable case of two classes (diamonds and stars). The blue line represents the case where the margin is maximum. A subset of samples determines the location of the decision boundary (represented by the red lines). Samples at (2,2) and (5,5) do not influence the decision boundary in this case and are not support vectors.

product $\mathbf{w}.\mathbf{u}$. Our objective is to determine which side of the decision boundary lies the data point, which can then reveal its class membership. To classify the data point as a positive sample we require that $f(\mathbf{u}) = \mathbf{w}.\mathbf{u} + b > 0$, for which the positive and negative samples can be constrained as

$$f(x_+) = \boldsymbol{w}.x_+ + b = +1 \tag{5.14}$$

$$f(x_-) = \boldsymbol{w}.x_- + b = -1 \tag{5.15}$$

Subtracting Equation 5.15 from Equation 5.14, we obtain $\mathbf{w}(x_+ - x_-) = 2$. Dividing by $\|\mathbf{w}\|$, the width of the margin can be computed to be

$$\frac{\mathbf{w}}{\|\mathbf{w}\|}(x_+ - x_-) = \frac{2}{\|\mathbf{w}\|} \tag{5.16}$$

This implies that in order to obtain the maximum margin for the SVM, we need to maximise $\frac{2}{\|\mathbf{w}\|}$. This is equivalent to minimising $\|\mathbf{w}\|^2$, and the optimisation problem becomes

$$arg \min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{5.17}$$

where $\frac{1}{2}$ is added for mathematical convenience. Bishop [131] describes the use of Langrange multipliers as a potential solution. It transforms a constrained optimisation problem into a convex optimisation problem.

It needs to be remembered that the minimisation is still constrained. The problem can be written as $y_i(\mathbf{w}x_i + b) \geq 1$ where y_i is +1 for a positive sample and -1 for a negative sample. In practice, it is unrealistic to expect to build a hyperplane that perfectly separates different classes. This can be accounted for by the introduction of a slack parameter (ξ) that offers flexibility while defining the hyperplane. Cortes and Vapnik [133] rephrased the minimisation problem as:

$$\min\|\boldsymbol{w}\| \, subject \, to \begin{cases} y_i(\mathbf{w}.x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0; \sum \xi_i \leq C \end{cases} \tag{5.18}$$

where the capacity parameter C determines the degree of penalisation for misclassifications. A large value of C will result in overfitting a decision boundary. Therefore the balance between generalisation and accuracy should be considered carefully while tuning an SVM. The slack parameter (ξ_i) imposes a non-zero cost for each x_i since the soft margin formulation is to allow for misclassified points (e.g. noisy points and/or outliers) at a cost.

The kernel trick - The above formulation of SVMs begins to break down when the class samples not linearly separable. To circumvent this scenario, a 'trick' may be employed which relies on the notion that data can be linearly separated if the features are transformed to a high dimensional space. Solving the optimisation problem reveals that the maximisation of

the margin depends only on the dot product between pairs of points (x_i and x_j). A kernel function can hence perform feature transformation:

$$K(x_i;x_j) = \phi(x_i)\phi(x_j) \tag{5.19}$$

Instead of resorting to a computationally expensive inner product ($\phi(x_i)\phi(x_j)$), the kernel function simplifies feature transformation. In other words, the kernel function helps to find the separating hyperplane in the feature space without the need to explicitly represent the space. Common kernel functions include the polynomial kernel, linear kernel, and the Radial Basis Function (RBF) kernel.

SVMs have shown excellent performance in different pattern recognition applications [134]. For our task of identifying the subpopulation of patients who will have different CVD risk scores by the HI and LI WHO/ISH charts, we used a linear kernel SVM, defined by $k(x_i,x_j) = (x_i.x_j + 1)^d$. For the case of the linear kernel, the degree parameter d=1. Sufficient accuracy was obtained when the linear kernel was used. A non-linear kernel (RBF) was also tested but provided similar performance and therefore the linear kernel was adopted in the final model. The trade-off between model complexity and classification error was achieved by optimising C through a grid search of the feature space using the training dataset. A suitable value of C was obtained by optimising on the area under the curve (AUC) parameter, also known as the C statistic. An internal four-fold cross validation (a technique described in the next subsection 5.2.4) was employed to optimise the hyperparameter C through a grid search. The proportion of class imbalance was consistent across the folds. An average of the C across the folds was taken and the SVM was retrained. LIBSVM, the SVM implementation used in this analysis, provides a 'weight' parameter w that was set to be inversely proportional to the size of samples in each class [135]. The penalty term C was multiplied by w for each class i to account for the imbalance. An alternative approach to examine the effect

of class imbalance on the SVM was also tested. The SVM was trained with the following class proportions: 50% of the training data comprising 70% samples of class 1 (n_{atr1}=108); the other 50% of the training data comprising a random selection of samples from class 0 (n_{atr1}=108). Total training data in this alternative approach was 216 samples and the rest of the samples were test data (n_{ate}=850). The training and testing of this alternative approach was repeated 10 times and the average test performance of the SVM was examined. The AUC evaluation metric is suitable and not influenced by class skewness (defined as the ratio of negative examples to positive examples) as opposed to other evaluation metrics like accuracy or Cohen's Kappa [136]. The libSVM implementation by Chang and Lin [135] was used in the Matlab programming environment.

Random Forests

A Random Forest (RF) is an ensemble classification technique that is broadly derived from the CART (Classification and Regression Tree) algorithm family [137]. It uses bagging (bootstrap aggregation) where N subjects or patients are sampled but with replacement, and selects a random subset of features, the number of which is $mtry$, to build decision trees. The decision trees in a RF are un-pruned [138] since the potential for overfitting is reduced, because (a) an ensemble of trees is utilised and (b) individual trees are bootstrapped. The trees in a RF are usually grown to the largest extent possible which could help learn irregular patterns in the dataset. A tree is grown until homogeneity is reached, which can be measured through the *Gini impurity* criterion.

$$I_G(e) = 1 - \sum_k p(k|e)^2 \tag{5.20}$$

where $I_G(e)$ is the Gini impurity at node e, $p(k|e)$ is the probability of class $k \in [0,1]$ at node e. For the binary classification task, Equation 5.20 reduces to $I_G(e) = 2p_0(1-p_0)$, where p_0 is the probability for class 0. When the number of samples having class 0 and class 1 are

equal, the impurity is maximum with I_G=0.5. When I_G=0, there is perfect purity. The results from each tree are aggregated in the end to perform classification or regression. Breiman [138] suggested two parameters that influence the generalization error of an RF: the *strength* of each tree and *correlation* between trees. With an increase in $mtry$, both the correlation and the strength of trees were found to increase [138]. However, an increase in the correlation leads to an increase in error rate, while an increase in the strength between trees reduces the error rate. Therefore $mtry$ needs to be optimised and is one of the two tuning parameters; the other being the optimal number of trees $ntree$ needed for achieving high accuracy.

An interesting component of RF is that of Out-of-bag (OOB) samples. During the construction of each tree, about one-third of samples from the original dataset are left out. The aggregation of these unused samples is called the OOB samples. They provide an unbiased out-of-sample or test data. Furthermore, OOB samples can be used to tune the RF parameters $mtry$ and $ntree$.

An estimate of variable importance can also be obtained from the OOB samples. For a given feature $p \in \{1,...,m\}$, the OOB samples in each tree are utilised to permute x_p alone whilst keeping all other features unaltered. The resulting difference between the number of votes for the target class (class 1) derived from the classifier on the permuted OOB data versus the OOB data is the importance measure for feature x_p. Since this method relies on the extent to which permutation results in a decrease in the accuracy of the model, it is referred to as the *variable importance due to mean decrease in accuracy* method. Subsequently, it follows that features that do result in a large decrease in mean accuracy as a result of perturbation are more important to the model as opposed to those features that have no effect on model performance due to perturbation [139].

RFs have numerous advantages including efficient handling of large data, robustness to outliers, fewer tuning parameters, automatic determination of variable importance, high classification accuracy, and methods to handle missing data [138]. We used the Matlab implementation of the randomforest R package [140].

5.2.4 Model validation

Generalisation of model performance - A popular method to obtain a reliable estimate of the test error and classifier performance on unseen data is the use of cross validation. Data from the training set is split into K folds. For each fold $k \in \{1, ..., K\}$, a model is trained on all folds except the k^{th} fold. The model is then tested on the k^{th} fold. This process is repeated K times. Subsequently, the performance is averaged across all the folds and a forecast of the test error is made. Each sample in the dataset acts as a test data point only once yet is utilised in training data $K-1$ times [141].

We performed four-fold cross validation on the training set to optimise hyperparameters and obtain the best performing model. For RFs, the OOB samples in the training set were used to optimise the number of trees and number of variables sampled for each split. The OOB samples acted as an internal test set for each tree. The resulting error estimate has been proven to be unbiased [142]. For regularised LR, the mean of the regularisation parameter was derived through four-fold cross validation. This was used to re-train the classifier and performance on all of the training set was estimated. The final model performance for all classifiers was evaluated on the test set.

5.2.5 Evaluation Metrics

To evaluate the models, we use the area under curve (AUC) or C statistic, which is a measure of the classifier's discriminative ability. We also used the F-score to select thresholds from the Receiver Operating Characteristics (ROC) curve. The F-score is the harmonic mean of the precision and recall, as defined below.

Recall

The recall or sensitivity (Se) is the true positive rate, or the number of times a patient belonging to the subpopulation is correctly identified. This means the number of times a patient for whom a TC test is beneficial for CVD risk prediction by the WHO/ISH models is correctly identified. It is defined as

$$Recall = \frac{TP}{TP+TN} \tag{5.21}$$

where TP indicates true positives and TN indicates true negatives.

Specificity

The specificity (Sp) is the true negative rate, or the number of times a patient who does not belong to the subpopulation is correctly identified. It is defined as

$$Specificity = \frac{TN}{TP+TN} \tag{5.22}$$

where TN indicates true negatives and TP indicates true positives.

Precision

The precision or positive predictive value is the proportion of true positives amongst all positive results and is defined as

$$Precision = \frac{TP}{TP+FP} \tag{5.23}$$

where FP indicates the number of false positives.

F-score

The generalised F-measure is given by

$$F_\beta = (1+\beta^2).\frac{precision.recall}{\beta^2.precision + recall} \tag{5.24}$$

where β measures the effectiveness of retrieval with respect to someone who attaches β times as much importance to recall as precision. We report the F_1, F_2, and F_3 scores where $\beta = 1$, 2, and 3 respectively.

The APRHI project was approved by the Institutional Ethics Committee of the CARE Hospital, Hyderabad in India and the University of Sydney Human Research Ethics Committee, New South Wales, Sydney. Participants provided informed, written content to contribute data to the study. The analysis carried out in this thesis chapter used de-identified, anonymised patient data from the APHRI project.

5.3 Results

5.3.1 Objective 1: Relative performance of LI and HI WHO/ISH CVD risk prediction charts

Out of 1066 subjects in *Dataset-1*, the LI and HI risk prediction charts misclassified 155 subjects (or 14.5%) relative to each other as shown in Figure 5.3. Hence the subpopulation consists of 155 subjects. Statistical significance testing was performed using the non-parametric Friedman's test, which is suitable for ordinal data (since the WHO/ISH charts predict in quantised ranges). The choice of WHO/ISH risk prediction chart was statistically

significant (p=0.008;χ^2=7.03).

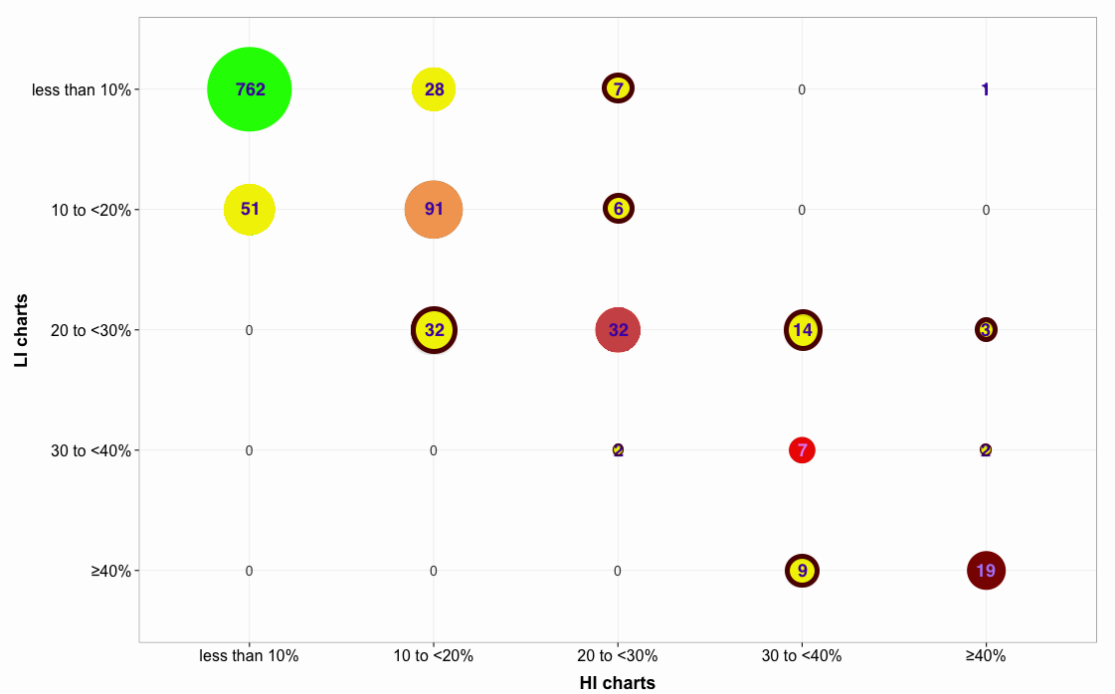


Figure 5.3 CVD risk prediction using the LI and HI WHO/ISH risk prediction charts on the chosen subset of APHRI data (N=1066). The five CVD risk ranges for whom the predicted risk using LI and HI WHO/ISH risk prediction charts concurred, are colour coded as follows: less than 10% risk in green; 10 to <20% risk in orange; 20 to <30% risk in red; 30 to <40% in light red; and ≥40% risk in maroon. Subjects for whom the predictions differed are highlighted in yellow circles with those that can be T_{HR} having a red highlight.

Figure 5.4 illustrates through a Venn diagram the number of subjects classified to be T_{HR} (or high-risk subjects requiring treatment) by both the LI and HI WHO/ISH models. The predictions from LI model resulted in 108 subjects being classified as T_{HR} while the HI model classified 88 subjects to be T_{HR} upon application of the previously described NPCDCS clinical guidelines. Both LI and HI models concurred on 80 subjects to be T_{HR} (p=0.2568 using Friedman's test, implying the choice of prediction model is not significant). However, they disagreed on 36 subjects overall (p<0.01 using Friedman's test, implying choice of prediction model is significant), which is 31% of all subjects identified as T_{HR} by

either the HI or the LI model. If it can be assumed that the HI model is more accurate on account of the additional information used in the prediction, we observe that 28 T_{HR} subjects are misclassified. This implies that the LI model over-predicts high-risk subjects requiring treatment as compared to the HI model.

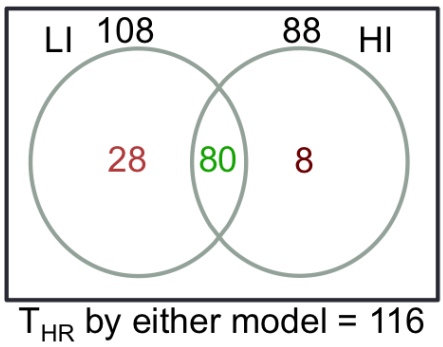


Figure 5.4 Venn diagram illustrating the number of subjects classified to be T_{HR} when the LI model and HI models of the WHO/ISH CVD risk prediction charts are used.

Table 5.1 shows the statistical characteristics of the misclassified subjects. We can observe the levels of certain mean risk factors that are at the thresholds of what is deemed normal and abnormal as per clinical guidelines. For example, the cut-off for diagnosis of hypertension is BP> 140/90mmHg and it is observed that the SBP in Table 5.1 for both males and females is on the borderline. This indicates that those subjects who will be the subject of step thresholds (as opposed to a smooth change) are prone to misclassification.

5.3.2 Objective 2: Models to identify subjects who benefit from TC testing

For the task of identifying the subpopulation who would benefit from TC testing, three classification approaches namely RLR, SVM, and RF were used (as described previously in Section 5.2.3). The training and out-of-sample performance of the classifiers are summarised

Table 5.1 Statistical characteristics of the subpopulation (n=155), which comprises those patients for whom the predictions for LI and HI WHO/ISH risk prediction charts differ.

Feature	**Male**	**Female**
n	57.4% (89)	42.6% (66)
Age, mean±sd, (years)	62.1±9.4	62.5±9.1
Current Smoker,%(n)	53.9% (48)	6.1% (4)
Systolic BP, mean±sd, (mmHg)	140.7±23.4	140.8±21.6
Diastolic BP, mean±sd, (mmHg)	81.7±13.2	78.8±12.3
Glucose, mean±sd, (mg/dl)	108.2±41.4	109.9±52.2
Total Cholesterol, mean±sd, (mg/dl)	187.4±57.5	206.1±61.0
Treated for hypertension,%(n)	29.2% (26)	34.8% (23)

in Table 5.2. RF has the highest AUC of 0.85 while the performance of SVMs is comparable (AUC 0.84). The ROC curve for the training data is shown in Figure 5.5. The F_1, F_2, and F_3 scores were computed and Figure 5.5 shows thresholds where the scores were maximal on the training data. Subsequently, the same thresholds were applied on the test data and the resulting sensitivity and specificity are described in Table 5.2. For RF, the OOB samples during training were used to calculate the F-scores since the AUC on the training data is perfect (AUC 1.00). The F_1 scores offer a balance between sensitivity and specificity, while F_3 scores emphasise sensitivity over specificity. For example, the performance of SVMs on test data shows that at the maximum F_1 score, a sensitivity of 91% and specificity of 69% was achieved while at the maximum F_3 score, we obtain 96% sensitivity and 66% specificity. The alternative approach to test the effect of class imbalance on the SVM (outlined in Section 5.2.3) was performed. An out-of-sample AUC of 0.83±0.017 was achieved. This is comparable to the standard approach of weighting the classes.

Table 5.2 Discriminative ability of Support Vector Machine (SVM), Random Forest (RF), and L1-Regularised Logistic Regression (RLR) to classify patients likely to benefit from a TC test. The F_1, F_2, and F_3 scores were obtained from the training data and used to threshold the out-of-sample test data (indicated by †).

Classifier	**C statistic**	**F_1 measure**			**F_2 measure**			**F_3 measure**		
		F_1 score	Se (%)	Sp (%)	F_2 Score	Se (%)	Sp (%)	F_3 Score	Se (%)	Sp (%)
SVM	0.87	0.54	90	76	0.72	95	72	0.82	95	72
SVM†	0.84		91	69		96	66		96	66
RF	0.84	0.51	82	76	0.68	91	70	0.78	93	67
RF†	0.85		87	71		91	66		91	62
RLR	0.86	0.54	81	80	0.69	86	77	0.79	96	62
RLR†	0.82		75	74		81	71		98	56

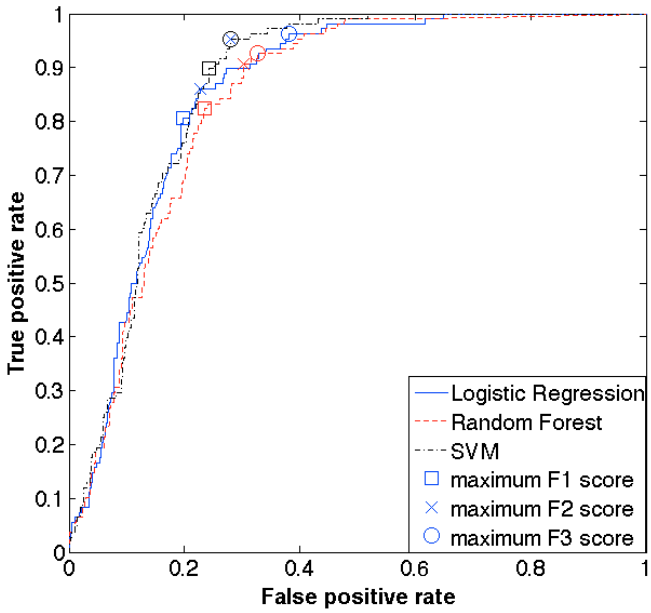


Figure 5.5 Receiver Operating Characteristic Curves for the SVM, RF, and RLR models for training data. The out-of-bag samples during training were used to obtain the ROC curve for RF since the performance on the training data was perfect (1.00). The F1, F2, and F3 scores were computed for every point on the AUC of the training dataset and the thresholds where the scores were maximum are highlighted.

The regularised logistic regression model identified two features as non-redundant from the entire feature set, namely age and SBP. The coefficients of the RLR model are given in Equation A.1. A similar inference on the most predictive features can be drawn based on the variable importance plot for the RF as shown in Figure 5.6. The importance plot shows the mean decrease in accuracy caused by a feature using the OOB samples. The larger the mean decrease in accuracy, the more important the feature is deemed to be - and according to Figure 5.6, age was most important followed by SBP.

$$\boxed{logit(c) = -5.6554 + 0.0416 * Age + 0.0132 * SBP} \tag{5.25}$$

where c is the probability that a patient requires a cholesterol test. c can be thresholded using the F1 or F3 score depending on the desired emphasis between recall and precision one wishes to have. Appendix A illustrates the use of this POC approach through worked examples.

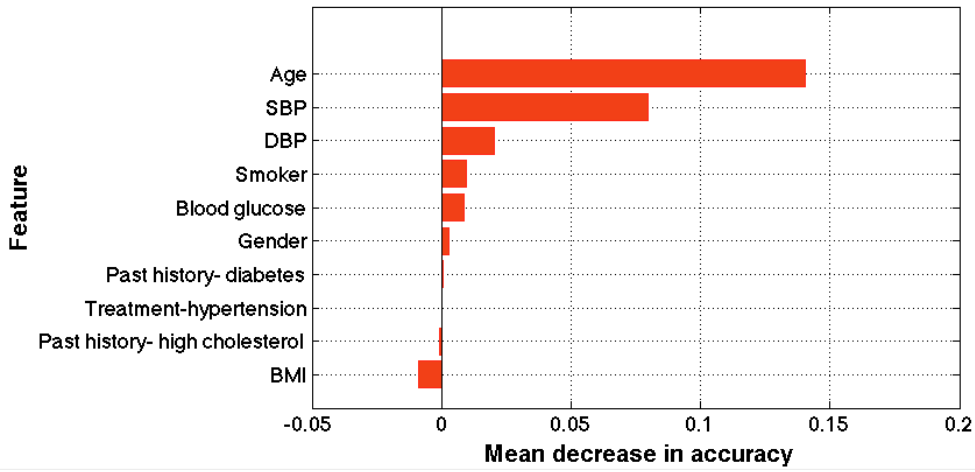


Figure 5.6 Variable importance ranked according to the mean decrease in accuracy (see Section 5.2.3) using the RF OOB samples. The permutation of age and SBP from the OOB samples resulted in the largest decrease in accuracy.

5.4 Discussion

We have shown that in a rural Andhra Pradesh population, the choice of version of WHO/ISH algorithm is important because there is a 14.5% difference in the predicted risk ranges between the LI and HI CVD risk charts. We also observed that the mean values of risk factors (especially SBP) in Table 5.1 are higher as compared to other risk factors in the chosen APHRI population (Table 4.4). In particular, SBP is close to the cut-off for what is deemed normal and abnormal as per clinical guidelines. This indicates that it is important to accurately stratify the risk in order to avoid missing out or under-treating people who may be at a higher risk and vice versa. The statistical approaches presented here can offer population-specific insights to resource allocation and cost saving. In countries like India, the cost of a cholesterol test has been reported to vary between $4 to $30 [143]. Although the Indian government's health spending has been increasing, the per capita expenditure is still only $61 per annum [144]. This implies that cholesterol testing has to be selective when population-wide strategies such as CVD risk screening are performed. Moreover, access to POC cholesterol testing is difficult and it is often the most expensive part of the CVD risk assessment.

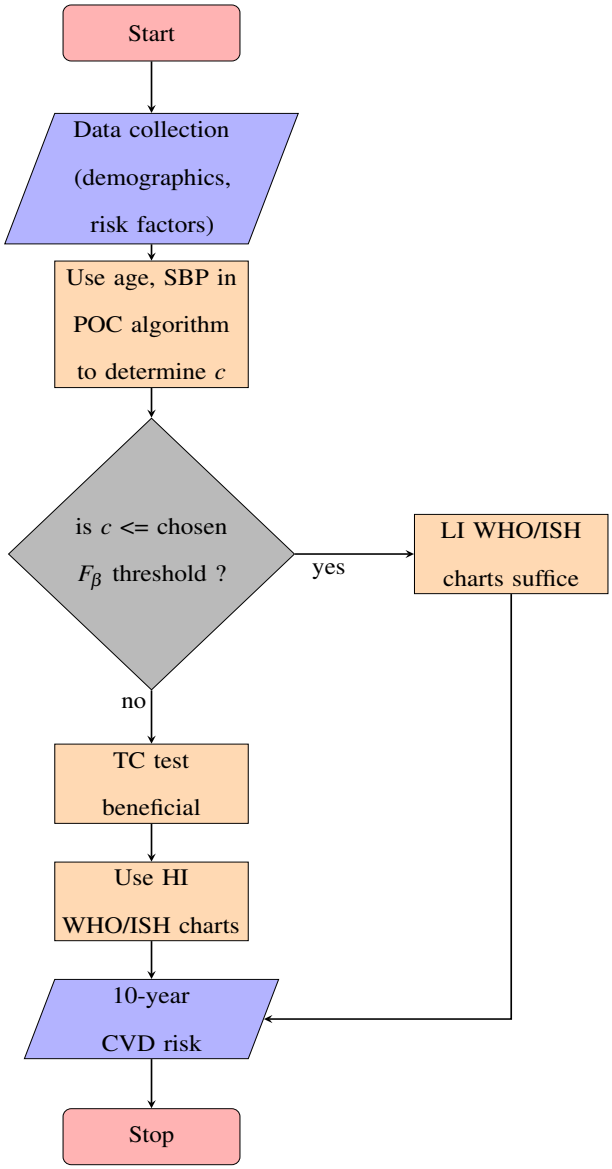
There can be substantial treatment implications for patients who are misclassified. If the HI model is assumed to be more accurate, screening using the LI chart will over predict risk for people at high CVD risk and we observed 28 subjects (which is 32% of those at T_{HR} according to HI model) to be misclassified as T_{HR}. This can have two major adverse implications in terms of over-treatment and prioritising patients for re-assessment/follow-up. We estimate the average cost of treating a high-risk patient is Indian Rupees (INR) 15 to 25 per day (or INR 450 to 650 per month). Out of 1.252 billion people in India [145], 368 million (29.4%) are aged over 40 years [146]. From the large-scale data collection described

in chapter 3, we can estimate at least 15% of those above the age of 40 years (55 million people) to be at high CVD risk. Therefore, in the worst case, we could have 18 million people who can be over-treated if the LI risk prediction charts were used, which would result in an expenditure of INR 8 to 13 billion a month. With a projected 61.5 million cases in India [147], with over 3 million new cases each year, our approach could potentially save substantial money in testing and tens of thousands of hours of human labour each year. Of course, all of this holds good if the WHO/ISH HI chart is more accurate, a premise which is still to be verified.

The final POC model comprises two features that is desirable in terms of the trade-off between model complexity, real life use cases, and accuracy, resulting in a *parsimonious* model. The variable importance plot (Figure 5.6) from the Random Forest classifier shows that Age and SBP are much more predictive than other features. Also, smoking status and gender, though predictive, are removed after L1 regularisation for the Logistic Regression model. This was performed to constrain the beta coefficients (otherwise susceptible to high variance) and reduce over-fitting. Furthermore, Age and SBP were consistently identified (after regularisation) as the only non-zero coefficients across 4-fold cross validation. This indicates that the inclusion of additional variables is likely to increase model complexity (or variance) at the expense of minor improvements in accuracy (bias reduction).

The advent of electronic tools such as tablets is an opportunity to package advanced statistical machine learning algorithms to assist risk predictions at point-of-care. We have demonstrated a pilot study (Chapter 2) as well as a large-scale baseline study (Chapter 3) that utilises a clinical decision support for minimally trained health workers through mobile devices for CVD risk prediction and management. Such a mobile based decision support system could be programmed to identify if the patient would benefit from a choles-

terol test before the CVD risk was computed. For instance, at point-of-care, the health worker could perform data collection on the patient's age (and other demographics) and blood pressure. The RLR algorithm presented here, which when installed on the mobile device, could then use the age and SBP data to advise on whether a total cholesterol test was necessary prior to CVD risk prediction with the WHO/ISH risk prediction charts. This process for decision support can be understood further through the flowchart illustrated below.



One of the main limitations of the work presented here is the fact that the dataset was acquired from a southern India population which might not be representative of the whole of India because of the large genetic diversity [148] and socio-economic inequities. However, rural regions of India are possibly in different stages of epidemiological transition. For instance, BP levels in Andhra Pradesh are largely identical to those reported in urban India [149][150] for similar age groups, which suggests that the Andhra Pradesh region is at an advanced stage of transition. Chow et al. [116] concur with this view by suggesting that rural regions in India are likely to develop risk factors levels comparable to those recorded in the APHRI study.

The lack of recorded CVD outcomes in the APHRI study is a barrier to validating both the LI and HI models of the WHO/ISH risk prediction charts with a gold standard. However, it is worth noting that there have been few or no large-scale prospective studies with recorded CVD events in India, especially ones that also record lipid profiles [116]. Also, a large number of patients in *Dataset-1* were of low CVD risk according to the WHO/ISH risk prediction charts. The WHO/ISH CVD risk prediction charts do not discuss if the LI model or HI model is more accurate in predicting CVD risk. As has been mentioned, we have made the assumption that the HI model is more accurate since it requires more information for prediction, but this assumption does need to be validated.

5.5 Conclusion

In this chapter, we found that the choice of LI or HI WHO/ISH risk prediction charts (for SEAR-D) was statistically significant for CVD risk prediction in rural Indian residents ($p=0.008; \chi^2=7.03$). The LI and HI risk prediction charts misclassified 155 subjects (or 14.5%) relative to each other. In terms of clinical relevance, the LI and HI models disagree on the T_{HR} status of 36 patients (31% of all subjects identified as T_{HR} by either model). If

the HI model is assumed to be more accurate, the LI model is observed to overpredict CVD risk for T_{HR} patients. Our POC test leverages two patient-specific risk factors, namely age and SBP, that are collected anyway during risk assessment with the WHO/ISH charts. This could assess the benefit of total cholesterol testing before risk computation and subsequently pre-determine whether the LI model is preferential to the HI model. The analysis in this chapter found good discrimination of the POC test with out-of-sample AUCs of 0.85 (RF), 0.84 (SVM), and 0.82 (RLR). The performance of RF on test data shows that at the maximum F_1 score, a sensitivity of 87% and specificity of 71% was achieved while at the maximum F_3 score, we obtained 91% sensitivity and 62% specificity. An understanding of the differences in risk prediction between the LI and HI models, and adoption of a pre-screening POC test to assess the benefit of a TC test, can aid planning for resource-allocation and saving costs for large-scale screening programmes.

Chapter 6

Machine learning for CVD risk prediction using a benchmark dataset

6.1 Introduction

In the last chapter, the non-linear associations between gender and diabetes were observed in the WHO/ISH risk model. Machine learning methods can model non-linearities in the data and, if applied properly, provide good generalisation on previously unseen data. When dealing with datasets where the time to event is of interest, one generally needs to account for participants who either dropped out early or joined after study commencement or both.

Zupan et al. [151] proposed a weighting technique to handle censored data for machine learning methods. They used the Kaplan-Meier estimate to derive a distribution of outcomes for patients who had short follow-up times and did not experience an event. These patients had two input entries where the corresponding outcomes were weighted according to the probability of the event and non-event respectively over the entire follow-up period. Although they could achieve predictive accuracies close to the Cox model, they could not out-perform it. The work by Kattan [152] arrived at a similar conclusion. They compared Cox regression

with artificial neural networks and several machine learning techniques, which resulted in the former providing comparable or superior predictive performance. The authors, however, expressed their work as 'an enabling technology' and suggested further testing on larger datasets.

Ripley and Ripley reviewed the use of neural networks by comparing it against linear models in survival analysis using breast cancer and melanoma data [153]. They imposed differential costs on errors (e.g. imposing penalties twice as high on false negatives as compared to false positives) and found that non-linear methods possessed substantial advantages. The authors also described ways in which the time to event problem could be reformulated as a classification problem. In this chapter, we follow one of the suggested approaches, which is to consider events within a fixed time (10 years).

Vanneschi et al. compared machine learning methods for prediction of survival in breast cancer patients on gene expression data [154]. They found that Genetic programming performed well and suggested further investigation of this method for cancer patient classification using gene expression data. Recent work by Khosla et al. [155] suggests that different feature selection techniques combined with machine learning classifiers can be used to achieve better performance than Cox regression, as demonstrated by the authors for the problem of stroke prediction. However, their result has so far not been replicated in other datasets.

The aims of this chapter are three-fold. Firstly, we wish to identify and obtain a subset of highly predictive features that are similar to *the type of inputs* required by our mobile-based CDSS, *SMARThealth*. For example, demographic and basic physiological measurements such as height and weight were collected with the rationale of being of clinical importance.

Features similar to height and weight such as waist circumference, neck circumference, or hip circumference will also be acquired relatively easily. Secondly, the aim is to compare CVD risk prediction achieved by a standard machine learning approach with established feature selection techniques, and that obtained with the latest version of the Framingham risk score (*FRS-3*) proposed by D'Agostino et al. [88]. After performing feature selection and building a model to predict CVD, the final aim will be to evaluate the performance of the model on data acquired from Indian subjects and compared with the *FRS-3* risk score and the WHO/ISH CVD risk prediction charts.

6.2 Methods

6.2.1 Framework for model development

An algorithmic approach to modelling focuses on deriving a function that maps the data to a response $\mathbf{y} \leftarrow f(\mathbf{X})$. A typical two-class classification problem can be formulated in this context as the occurrence of a CVD event (represented by class **1**) or not (class **0**) within a specified time duration (for instance, 10 years). We represent the data through a 'design matrix' $\mathbf{X}$ containing n patients and m features, and the outcomes via vector $\mathbf{y} \in [0,1]$.

6.2.2 Data

Data from Exam 6 of the Framingham offspring cohort (or *Dataset-2* as detailed in Chapter 4) was considered for analysis. The offspring cohort was chosen for analysis as the 3rd generation cohort did not record enough events (65 out of 3391) while the original cohort dataset was collected in 1958, and apart from the lack of heterogeneity in the dataset (on account of diversity), the availability of parameters like cholesterol was limited during early Exam cycles. The choice behind the particular Exam cycle is attributed to two major reasons: firstly because Exam cycle 6 recorded higher event rates (11.35%); and secondly

because data on family and medical history of the patients were available (which constitute inputs required in the mobile-based CDSS). Moreover, Exam cycle 6 consists of a large list of features that are similar to the type of inputs required by our mobile-based CDSS, *SMARThealth*. The task at hand is to identify and obtain a subset of highly predictive features.

Exam cycle 6 consisted of 3434 patients out of whom 3040 were free of CVD. 519 variables were recorded that included data on pulmonary function tests and respiratory surveys. The outcome was whether CVD was developed within 10 years from Exam cycle 6. After feature selection and development of a prediction model for estimating 10-year CVD risk, the resulting model was applied to data from rural India (*Dataset-1*) comprising 1066 patients. The summary of the methodology followed in this chapter is shown in Figure 6.1.

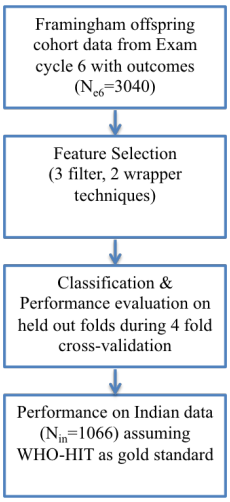


Figure 6.1 Overview of the work presented in this chapter.

6.2.3 Pre-processing

Missing data was imputed using the K-nearest neighbour (KNN) technique. KNN has been shown to be effective in imputing both discrete values (through selection of the most frequent value in k-nearest neighbours) and continuous values (through the mean of values in k-nearest

neighbours) [156]. For simplicity and computational efficiency, k was chosen to be 3. Four features that had more than 40% of the constituent values missing were removed. The dataset then comprised 515 variables, out of which 149 features that were similar to the type of input parameters from Steps 1-4 of the mHealth CDSS were initially chosen to build the design matrix. This included the family of features under the following categories: family history, medication history, alcohol consumption, anthropometry, and clinical diagnostic impressions (CDI). Examples of features selected included waist-to-hip ratio (anthropometry), consumption of units of beer/week (alcohol consumption), cigarette type (smoking status), calcium channel blockers (CV medication), degenerative joint disease (CDI).

A further 16 features were incorporated into the design matrix including the mean and standard deviation of all recorded BP measurements; ratios of total cholesterol to HDL, and waist to hip; combination of medical history of parents (for heart attack, stroke, high blood pressure, hyperlipidemia); and computation of elevated blood glucose status as per the American Diabetes Association guidelines [63]. Certain binary features in the Framingham dataset also encompassed an 'unsure' option (and hence three categories - yes, no, and unsure). Such features were converted to binary values and had an extra 1 assigned where the option was 'unsure' and 'zero' otherwise (e.g. smoker = 1,0, non-smoker = 0,0, unsure = x,1 where x is treated as missing data and imputed as described above).

The design matrix $\mathbf{X}$ was normalised using the Z-score normalisation (Equation 6.1), which for the j^{th} feature is given by

$$\mathbf{x}_j^* = \frac{\mathbf{x}_j - \mu_j}{\sigma_j} \tag{6.1}$$

where $\mathbf{x}_j^*$ is the normalised value of $\mathbf{x}_j$, μ_j and σ_j are the mean and standard deviation for the j^{th} feature, respectively.

6.2.4 Feature selection

The concept of feature selection was introduced in the previous chapter (Section 5.2.2). Building a prediction model with large number of features (such as the 519 features in the dataset) may not be optimal because of the curse of dimensionality. A reduced feature set offers more clarity on the covariates through analysis of the most predictive features [157]. In order to achieve a reduced feature space, we may resort to either feature transformation or feature selection. Feature transformation involves combining features linearly or non-linearly to a create a new feature space that is a condensed representation of a few or all of the original features. A subset of the new feature space may then be chosen for classification. Examples of feature transformation include Principal Component Analysis (PCA) and factor analysis.

Feature selection, on the other hand, reduces the feature space through selection of an optimal subset of predictive features from the original features. That is, given m features, the job of a feature selection technique is to pick a subset m_{red} such that $m_{red} < m$. Feature selection techniques may be largely classified into two types - Wrappers and Filters, and sometimes a third category called Embedded techniques is also used. Wrappers incorporate a learner to perform subset selection while filters tend to extract the information content using common statistical metrics such as correlation or mutual information. An example of a filter method is the maximum relevance minimum redundancy approach [122] which was described in the previous chapter. mRMR only uses an information measure, and is defined as

$$mRMR = \max_{S} \left\{ \frac{1}{|S|} \sum_{\mathbf{f}_i \in S} MI(\mathbf{f}_i; \mathbf{c}) - \frac{1}{|S|^2} \sum_{\mathbf{f}_i, \mathbf{f}_j \in S} MI(\mathbf{f}_i; \mathbf{f}_j) \right\} \tag{6.2}$$

where S is the feature set, I is the mutual information between feature $\mathbf{f}_i$ and class $\mathbf{c}$ or between features $\mathbf{f}_i$ and $\mathbf{f}_j$.

Embedded methods use a classifier as part of the feature selection process (e.g. forward selection with an ensemble of decision trees) and most often combine filter and wrapper methods. In this chapter, four other widely used filter and wrapper methods (namely the Gram-Schmidt Orthogonalisation (GSO), RELIEF criterion, Elastic Net, and sparse Linear Discriminant Analysis) were chosen as feature selection techniques in addition to mRMR. The techniques are advantageous on account of their simplicity and effectiveness [157].

Gram-Schmidt orthogonalisation

GSO is a sequential forward feature selection technique [158]. The set of selected features is empty initially with the error being maximum. At every step, the feature that results in the greatest decrease in the error is added, and this process is repeated until the significant decrease in error is not observed with feature addition. To perform this, the feature that describes the outcome the most is selected by computation of the smallest angle of that feature with the outcome vector in N dimensional space of the given samples. This is achieved by

$$Cos^2\ (\mathbf{x}_i, \mathbf{y}) = \frac{(\mathbf{x}_i\ .\ \mathbf{y})^2}{\|\mathbf{x}_i\|^2 \|\mathbf{y}\|^2} \tag{6.3}$$

where $\mathbf{x}_i$ is the vector for feature i, $\mathbf{y}$ is the outcome vector, $\|\mathbf{x}_i\|^2$ is the square of the Euclidean norm of $\mathbf{x}_i$, and $\|\mathbf{y}\|^2$ is the square of the Euclidean norm of $\mathbf{y}$. The Euclidean norm (or L2 norm) is used since it represents the notion of the length of a vector (in this case the magnitude of each feature). The information that overlaps between the outcome and the remaining features is discarded subsequently through their projection on to a null subspace of the selected feature, and is given by $\mathbf{x}_{i,new} = \mathbf{x}_i - \mathbf{proj_{w1}}(\mathbf{x_i})$ where $w1$ is the space spanned by $\mathbf{y}$ and $\mathbf{proj}(\mathbf{x_i})$. GSO is conceptually similar to mRMR in the sense that at every step, the feature added is most correlated with the outcome and least correlated with all other features. Using an orthogonal transformation ensures that features are decorrelated which implies that each feature can be evaluated independently.

RELIEF

This is a feature weighting technique that is based on choosing k closest neighbours that are randomly selected for a Nearest Hit (NH) or Nearest Miss (NM). Let us consider this in the case of feature selection for a binary classification problem. Upon random selection of a sample x belonging to dataset $\mathbf{X}$, a same-class instance that is closest (as measured by Euclidean distance) is labelled as the nearest hit while an instance of different class that is closest is called a nearest miss. A weight vector that is updated for every randomly sampled case, is given by

$$\mathbf{W}_i = \mathbf{W}_i - (\mathbf{x}_i - \mathrm{NH}_i)^2 + (\mathbf{x}_i - \mathrm{NM}_i)^2 \tag{6.4}$$

A feature's weight is increased whenever it differs much more for k nearest instances of samples from the other class (e.g. class 0) in comparison to k nearest instances of the same class (e.g. class 1). After m iterations, we obtain the relevance metric given by $\mathbf{R} = \mathbf{W}/m$. The key advantages of RELIEF are its robustness to feature interactions, usage of few heuristics, high noise-tolerance and computational efficiency. The main disadvantage is that it generates a measure of relevance without taking into account redundancy [159].

Elastic net

Zou and Hastie [160] proposed a regularization method for the class of generalised linear models (GLM) [161] that linearly combines the L1 and L2 penalties in the LASSO and RIDGE regression [132] techniques, respectively. They defined a minimisation problem of the β coefficients with λ_1 and λ_2 penalties as

$$\hat{\beta} = argmin\ |\mathbf{y} - \beta\mathbf{X}|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1 \tag{6.5}$$

where $|\beta|^2 = \sum_{j=1}^{m} \beta_j^2$ and $|\beta|_1 = \sum_{j=1}^{m} |\beta_j|$. In practice, the terms λ_1 and λ_2 are combined to form a parameter $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1}$. When α is introduced in 6.5, the optimisation problem

becomes

$$\hat{\beta} = argmin\ |\mathbf{y} - \beta\mathbf{X}|^2, \text{constrained by} \alpha|\beta|^2 + (1-\alpha)|\beta|_1 \leq t; \text{for a defined t} \qquad (6.6)$$

It may be observed that when $\alpha = 0$, the equation reduces to the LASSO technique [132]. Whilst LASSO is excellent for building sparse models [162][160], it has two main limitations: first, when $m > N$, LASSO selects m variables at most; and second, from amongst a group of correlated variables, LASSO tends to select one variable and ignore the others [160]. When $\alpha = 0.5$ as in the Elastic net technique, the effects of both L1 and L2 penalties are interspersed equally. While the L1 penalty acts in favour of a sparse model, the L2 penalty encourages a *grouping effect* thereby removing restrictions on the number of selected features. This may be visualised in 6.2.

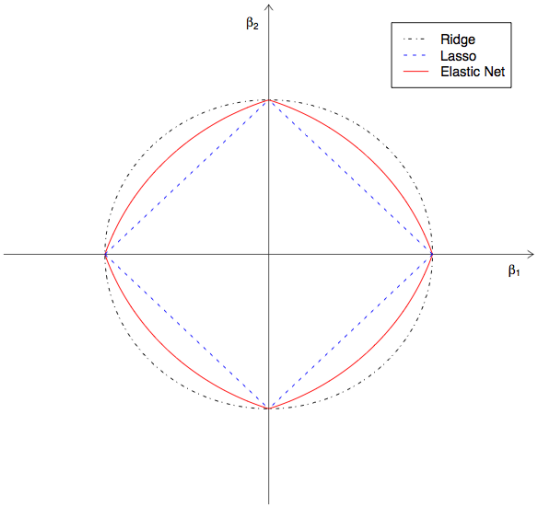


Figure 6.2 Illustrating the contours of LASSO, RIDGE, and Elastic Net regularisation techniques. Elastic Net results in convex edges and singularities at vertices, with the degree of convexity specified by α.

Sparse Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a method of classification that relies on maximising the within-class scatter and between-class scatter through projecting the data onto a

low-dimensional space. This is computed by obtaining the eigenvectors and eigenvalues of class-specific scatter matrices. A subset of eigenvectors with large eigenvalues are chosen to derive a transformation $\mathbf{w}$, and therefore $\mathbf{X}_{new} = \mathbf{w}^T.\mathbf{X}$. Secondly, an optimal $\mathbf{w}$ that maximises the separation between scatter matrices is computed by an objective function that defines the distance between projected sample means as

$$J(\mathbf{w}) = \frac{|\bar{\mu_1} - \bar{\mu_2}|^2}{\bar{\sigma_1}^2 + \bar{\sigma_2}^2} \tag{6.7}$$

where $\bar{\mu_1}$ and $\bar{\mu_2}$ are multidimensional sample means for classes 1 and 2, while $\bar{\sigma}_1$ and $\bar{\sigma}_2$ are the respective standard deviations. To maximise the objective function, $J(\mathbf{w})$ can be expressed in terms of w as

$$J(\mathbf{w}) = \frac{w^T.S_B.w}{w^T.S_W.w} \tag{6.8}$$

where S_W is the within-class scatter matrix and $S_W = S_1 + S_2$ and $S_i = \sum_{x \in w_i}(x - \mu_i)(x - \mu_i)^T$; S_B is the between-class scatter matrix given by $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$.

$$\frac{d}{dw}J(\mathbf{w}) = \frac{d}{dw}\frac{w^T.S_B.w}{w^T.S_W.w} = 0 \tag{6.9}$$

The solution for the above equation is $S_W^{-1}.S_B.w - J.w = 0$ which implies $w = S_W^{-1}(\mu_1 - \mu_2)$. Clemmensen et al. [163] imposed an elastic net penalty thereby enabling feature selection and classification with a sparse LDA (sLDA) classifier. LDA does not however, perform well if separation between the classes requires a non-linear boundary or if the discriminatory information is not encoded in the sample mean but in the variance [163]. A central assumption in LDA is that of normally distributed data with a common covariance matrix between two classes [164].

The process of feature selection (explained in Figure 6.3) was performed using four-fold cross validation where the data split into a training set and held-out set. Three filter methods

1: **procedure** FEATURE SELECTION
2: Given data **X** of n patients and m features
3: **repeat**
4: **for** k = 1:K
5: Apply desired feature selection technique $FS(X)$ on all subjects except those in the k^{th} fold.
6: Obtain ranks or 1...l positions, where l is the number of positions of interest (or ranks) for each feature.
7: Obtain matrix $M_{k,l}$ representing feature indices across k folds and l positions.
8: **end for**
9: **until** $r \in$1,...,T
10: Compute the frequency of occurrence (FREQ) from the derived subset of the features U that make up top P positions across T repetitions.
11:

$$FREQ_{i,j} = \frac{1}{T}\sum_{r=1}^{T}\frac{1}{K}\sum_{k=1}^{K}C(i,j,k) \tag{6.10}$$

12: where $i \in (1,U)$, $j \in (1,P)$, $k \in (1,K)$ and
13:

$$C(i,j,k) = \begin{cases} 1, & M_{k,j} == i \\ 0, & \text{otherwise.} \end{cases} \tag{6.11}$$

14: **end procedure**

Figure 6.3 Process of feature selection

(mRMR, GSO, RELIEF) and two wrapper techniques (Elastic net, sLDA) were applied on the training set which had data from all folds except the k^{th} fold (see Section 5.2.4 of the previous Chapter 5 for description of k-fold cross validation). A list of features that were selected in the top 10 positions over T =10 repetitions of this process for each technique was computed.

6.2.5 Classification techniques for detecting CVD events

Two classification techniques were chosen - a standard baseline approach with a probabilistic framework for classifying binary outcomes namely Logistic regression (LR), and an ensemble classification technique, Random Forests (RF). RFs offer advantages such as robustness to outliers and missing data, fewer tuning parameters, automatic determination of feature

importance, and high classification accuracy [165]. For more detail on LR and RFs, see sections 5.2.3 and 5.2.3 from the previous chapter 5. To handle the dataset imbalance arising on account of the low occurrence of CVD events in 10 years (11.35%), the RF classifier was programmed such that the two classes were weighted in proportion to the event rate. This influences the predictions by penalising mis-classifications in the minority class more than the majority class by a factor proportional to the inverse of the event rate. In LR, class-imbalance has an effect on the intercept of the model alone [166]. This may be advantageous, as we shall explore in the discussions section, on how the model can be affected by different incident rates (which consequently determine the proportion of the two classes)

6.2.6 Evaluating model performance-Discrimination

The performance of a classifier is usually described in the literature using two characteristics - discrimination and calibration. The discriminative power is the classifier's ability to separate outcomes, such as those who experience a CVD event from those who do not. A probabilistic output can be generated from the classifiers (LR and RF), with a threshold selected to convert to binary outputs. This can be compared with the observed outcome to derive additional discriminative metrics such as the predictive accuracy of the classifier. In order to achieve this, we first compare the binary output to the observed outcome and subsequently classify the predicted value as a true positive (TP), a false positive (FP), a true negative (TN) or a false negative (FN).

A true positive occurs when a patient who has a CVD event is correctly classified. The true positive rate or sensitivity is the proportion of correctly identified subjects with CVD amongst all observed positive cases. A true negative is a correct rejection of subjects without CVD. A false positive, also known as a Type 1 error, is the incorrect identification of a patient

without CVD. A false negative or a Type 2 error, is the incorrect rejection of a patient with CVD. Specificity is the proportion of correctly rejected cases amongst all observed negative cases. The accuracy is the proportion of correctly identified CVD cases amongst all cases.

Area Under Receiver Operating Characteristic

Variation of the sensitivity against (1-specificity) over the continuum of thresholds generates a Receiver Operating Characteristic (ROC) plot. It indicates all possible combinations of sensitivity and specificity for a classifier from which appropriate thresholds or operating points may be selected according to the clinical need. For instance, diagnostic tests at the primary care level often prefer a higher sensitivity to specificity. This is because it is often better to tolerate a type 1 error (false positive) than a type 2 error (false negative).

A line defined by sensitivity and specificity of 0.50 connects (0,0) to (1,1) of the ROC plot. The discriminative power conveyed by this line is that of chance (or a random guess). The ideal classifier would generate an ROC curve towards the top left corner (0,1) of the ROC curve. Integration of the ROC curve results in the Area under the receiver operating characteristic (AUROC, or simply AUC), which measures the discriminative ability of a classifier under all thresholds. In CVD risk prediction literature, the AUROC is also referred to as the c-statistic and is the most widespread measure for reporting discrimination [81]. AUROC can also be interpreted as a measure of the probability that a patient who develops CVD has a higher risk score than the probability of patient who does not develop CVD, i.e. $P(y_{1i} > y_{0i})$.

Class-imbalance occurs when the proportion of samples from a certain class is much greater than the other (in case of a binary classification task). The AUROC is a measure that is less sensitive to class imbalance as opposed to commonly used metrics like accuracy, Cohen's Kappa etc [136].

6.2.7 Evaluating model performance-Calibration

A classifier's calibration is an assessment of the how close the estimated probabilities are to the observed outcomes. For instance, the subgroup of patients with high CVD risk should correspond to these with higher incidence of CVD.

Hosmer and Lemeshow [167] proposed statistical tests to measure the goodness-of-fit of a model. The procedure for computing the statistics involves the estimated or predicted probabilities from the model, and the observed events. Firstly, the estimated probabilities are ordered according to increasing risk, and grouped into deciles. A comparison is then performed between the observed outcomes and estimated probabilities for each decile, as given by

$$HL_{\hat{C}} = \sum_{j=1}^{D} \frac{(O_j - E_j)}{n_j p_j (1 - p_j)} \tag{6.12}$$

where D is the number of bins, which is equal to 10 when split into deciles, O_j is the observed number of events at bin j, E_j is the number of estimated events at bin j, p_j is the average of predicted probabilities for the positive class (that is, for a CVD event) in bin j, and n_j is the number of subject in bin j. A p-value can be obtained that assesses the calibration of the model. A statistically significant model indicates poor calibration. The $HL_{\hat{C}}$ metric relies on the premise that equal numbers of patients are distributed across deciles. A modification of the Hosmer-Lemeshow uses deciles of risk, i.e. 0-10%, 10-20%, ... and is known as the HLH statistic. Kramer et al. [168] demonstrated that the Hosmer-Lemeshow statistic, which was tested with 200 samples in the original paper, is extremely sensitive to the sample size when calibration is not perfect, which is practically the case in most models.

Log-Likelihood The Hosmer-Lemeshow statistic has come under criticism for its sensitivity to the dataset size [169]. An alternative to this is the log-likelihood score which indicates both calibration and discrimination. For a binary classification problem, the log-likelihood is

defined as follows

$$\mathscr{L}(\beta|x) = \sum_{i=1}^{N} y_i log(p_i) + (1-y_i) log(1-p_i) \tag{6.13}$$

$$= \sum_{i=1}^{N} y_i(\beta_0 + \sum_{j=1}^{K} \beta_j x_j) + \sum_{i=1}^{N} log(1-p_i) \tag{6.14}$$

Brier Score Another metric that can provide a measure of both calibration and discrimination is the Brier Score (Br). It is calculated as the squared difference between the expected and observed outcomes as follows

$$Br = \sum_{j=1}^{N} (\hat{y_j} - y_j) \tag{6.15}$$

Brier scores are easy to implement, interpret, and the lower the brier score, the better is the calibration and discrimination of a predictive model.

6.3 Results

6.3.1 Feature selection

Table 6.1 describes results from the feature selection process that was previously described in Figure 6.3. The 29 features that appeared at least 70% of the time within the top 10 ranking for any feature selection technique are highlighted in blue. 8 features that appeared 60% or higher for more than 1 feature selection technique and within the top 10 ranking, are illustrated using a red font. The thresholds employed were heuristic but in tune with the aim to not select too few features (high threshold) nor too many features (low threshold). The 8 features were *age*, *gender*, *glucose*, *high density lipoprotein (HDL)*, *SBP1*, *Mother-died of heart disease*, *waist-hip ratio*, *cardiovascular medication - calcium channel blockers*, *medication - oral hypoglycaemics*. It is interesting to observe that our feature selection

process selected waist-hip ratio instead of the traditional BMI. The waist-hip ratio has been reported as a key indicator for abdominal obesity [170] [171], which is strongly associated with the development of type-2 diabetes [172] and acute myocardial infarction [173].

Table 6.1 Features that appeared in the top 10 positions on application of the five feature selection techniques. Cells coloured in blue highlight those features that appeared 70% or more within the top 10 ranks of at least 1 feature selection technique. Features with a red font colour were those that appeared 60% or more within the top 10 ranks of at least 2 feature selection techniques. Terminology: HDL - high density lipoprotein; SBP - systolic blood pressure (different BP measurements are suffixed by the order of measurement, e.g. SBP1, SBP2, and SBP3); DBP - diastolic blood pressure; HBP - high blood pressure; hdDeath - death due to heart disease; CV - cardiovascular

	mRMR	GSO	RELIEF	Elastic Net	sLDA
Risk Factors					
Age	100	100		100	100
Gender		3		99	100
Glucose	93	83			100
TC-HDL Ratio	4	3			10
HDL		100		97	
Elevated BP	4				
Hypertension				82	
SBP1	100	65		16	
SBP2		22		10	
SBP3	94				
DBP1	100	2			
DBP2					75
DBP3					100
meanSBP	100	5		14	
stdDBP					26
BEER-NUMBER OF DAYS DRINK PER WEEK		35	60	2	
Family History					
FATHER-EVER HAVE HIGH BLOOD PRESSURE		2		7	
FATHER-EVER HAVE HIGH BLOOD CHOLESTEROL	8			18	1
FATHER-EVER HAVE DIABETES MELLITUS			33	1	
FATHER-DIE OF HEART DISEASE			52		
MOTHER-EVER HAVE HIGH BLOOD PRESSURE			6		
MOTHER-EVER HAVE DIABETES MELLITUS			4		
MOTHER-DIE OF HEART DISEASE		80	100	99	
Family history HBP			80		
Family history hdDeath			86	10	
Medical History - Clinical Diagnostic Impression					
CDI-RHEUMATIC HEART DISEASE		37			
CDI-OTHER PERIPHERAL VASCULAR DISEASE		38			
CDI-PARKINSONS DISEASE		12			
CDI-URINARY TRACT DISEASE			2		
CDI-PROSTATE DISEASE		45	100	18	
CDI-RENAL DISEASE		4	17		

Continued on next page

Table 6.1 – continued from previous page

	mRMR	GSO	RELIEF	Elastic Net	sLDA
CDI-EMPHYSEMA		2			
CDI-CHRONIC BRONCHITIS			25		
CDI-GOUT		27	6		
CDI-DEGENERATIVE JOINT DISEASE			100	22	
CDI-OTHER NON CV DIAGNOSIS			26		
Smoking					
SMOKED CIGARETTES REGULARLY IN LAST YEAR		1			
HOW MANY CIGARETTES SMOKED PER DAY		1			
CIGARETTE STRENGTH				18	
CIGARETTE TYPE		97		13	
CIGARETTE FILTER		1			
CIGARETTE LENGTH				41	
SPOUSE-PIPES/DAY AT HOME		10		7	
Current Smoker		1			
Anthropometry					
Neck-circumference	1			9	99
Waist-Girth	1	9		88	99
Hip-Girth					90
Thigh-Girth		5			
Knee-Height					100
Waist-Hip Ratio	100	3			100
Medication history					
TAKE ASPIRIN REGULARLY			20		
ASPIRIN FREQUENCY				14	
USUAL ASPIRIN DOSE FOR ABOVE			2		
CV MEDS - LONGER ACTING NITRATES		1			
CV MEDS - CALCIUM CHANNEL BLOCKERS	66	100	19	100	
CV MEDS - THIAZIDE DIURETICS			62		
CV MEDS - POTASSIUM SUPPLEMENTS		1			
CV MEDS - RENIN-ANGIOTENSIN BLOCKING			100		
CV MEDS - PERIPHERAL VASODILATORS	100				
CV MEDS - OTHER ANTI-HYPERTENSIVES		9			
CV MEDS - ANTIPLATELET	100				
TREATMENT FOR BLOOD PRESSURE				19	
MEDS - ANTI-CHOL -NIACIN/NICOTINIC ACID		1			
MEDS - ANTI-CHOL -STATINS			1	4	
MEDS - ANTIGOUT-COLCHICINE		6			
MEDS - THYROID EXTRACT	1	8			
MEDS - ORAL HYPOGLYCAEMICS		36	99	92	
MEDS - ANALGESIC-NARCOTICS		31			
MEDS - ANTI-PARKINSON DRUGS		2			

6.3.2 Model performance

A total of 3040 patients with the 8 features chosen from the feature selection process was utilised to build a prediction model (herein referred to as ‘Model 1’) using the RF and RLR

classifiers. A plot of the features included in Model 1 versus obtained AUCs that describe model performance is shown in Figure 6.4, which also illustrates performance on the held out fold from a 4 fold cross validation procedure on Exam 6 of Framingham Offspring cohort. The increase in AUC with addition of features can be observed. Also, an AUC close to 0.70 can be achieved with LR with only age and gender included as features. The LR classifier interestingly performs better than the more recent RF classifier. The RF technique does not perform as well as LR with fewer features. However, it is comparable when the number of features is increased. Although the performance of classifiers is data-dependent and RF has traditionally shown achieved performance than LR [174], standard baseline techniques such as LR have not always underperformed in comparison to RF [175]. The performance of LR in Model 1 when compared to the *FRS-3* simple (LI) and main (HI) versions in the held out data (Exam 6) is only marginally superior.

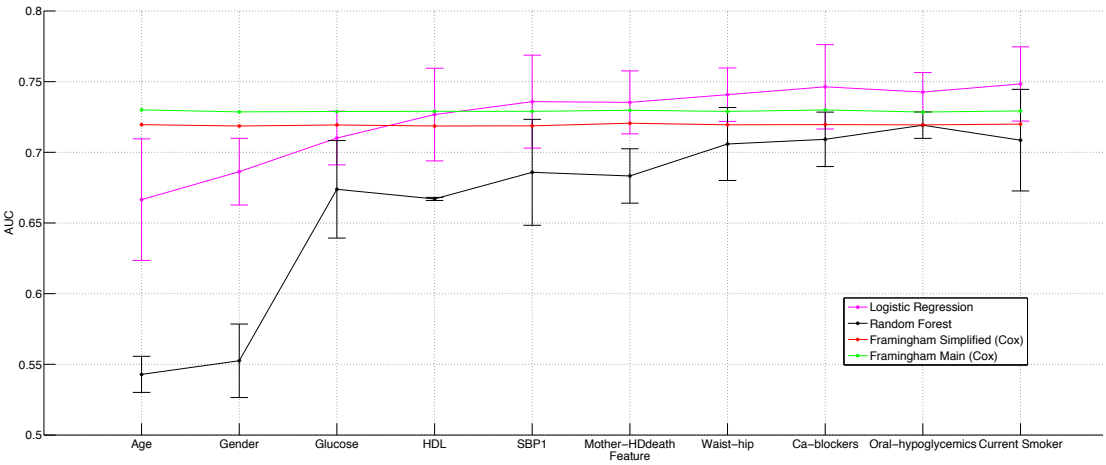


Figure 6.4 Comparison of AUCs of the *FRS-3* simple and main risk scores with LR and RF classifiers of Model 1. The performance on held out data during 4 fold cross validation is shown. Features on the x-axis are listed in a cumulative manner. The Framingham risk scores use different feature sets and their reported AUC has no corroboration with the features on the x-axis (and therefore is shown as a flat line).

Table 6.2 describes the performance of Model 1 and two versions of the *FRS-3* on held out data (Exam cycle 6).

Table 6.2 Classification results on the held out fold from a 4 fold cross validation procedure on data from Exam cycle 6. All 8 features that were selected were used in Model 1. Although Model 1 has higher AUC and higher log-likelihood scores in the held out folds, the performance is not substantially better than either version of *FRS-3*. The RF performance is inferior to the LR performance.

	Classifier	**Held out folds**		
		AUC	**Brier score**	**Log-Likelihood**
Model 1	*LR*	0.748 ± 0.026	0.093 ± 0.004	-239.994 ± 9.596
	RF	0.709 ± 0.036	0.097 ± 0.004	-251.314 ± 10.083
Framingham	$FRS-3_{simple}$	0.720 ± 0.009	0.097 ± 0.001	-251.704 ± 4.323
Risk Score	$FRS-3_{main}$	0.729 ± 0.013	0.095 ± 0.001	-247.043 ± 4.008

In Section 4.4 of Chapter 4, a comparison of the WHO/ISH charts and Framingham risk scores (*FRS-3*) was performed. Here the comparison is extended to include Model 1. The performances of the LR classifier of Model 1 and the two versions of the Framingham risk score (*FRS-3*), against the WHO/ISH high information (HI) risk prediction charts are shown in Figure 6.5. The risk prediction algorithms were then applied to the data from rural India (*Dataset-1*). Standard error metrics such as mean absolute error (MAE) and root mean squared error (RMSE) were used. MAE is computed as $\frac{1}{N}\sum_{t=1}^{N}|y_t - \hat{y}_t|$, while RMSE is computed as $\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$, where y_t is the WHO/ISH HI risk score and $\hat{y}_t$ is the risk score computed by Model 1 or *FRS-3* main or *FRS-3* simple. It is observed that the comparison between Model 1 and the WHO/ISH HI risk score resulted in the lowest MAE of 11.79 and RMSE of 13.59 when compared to *FRS-3* main versus WHO/ISH HI (MAE-13.62; RMSE-16.47) and *FRS-3* simple and WHO/ISH HI (MAE-13.49; RMSE-16.50). The differences in calibration can also be visually inspected in Figure 6.5. The scatter plot of a perfectly calibrated model would lie along the diagonal line which indicates the agreement of predicted probabilities of both models. It is clear from the plots (B and C) of either version of *FRS-3* against WHO/ISH HI, that the scatter points are aligned between 0 to 20% for both

FRS-3 scores irrespective of the WHO/ISH HI risk score. The scatter points in plot A of Model 1 versus WHO/ISH HI seem to align better along the diagonal line.

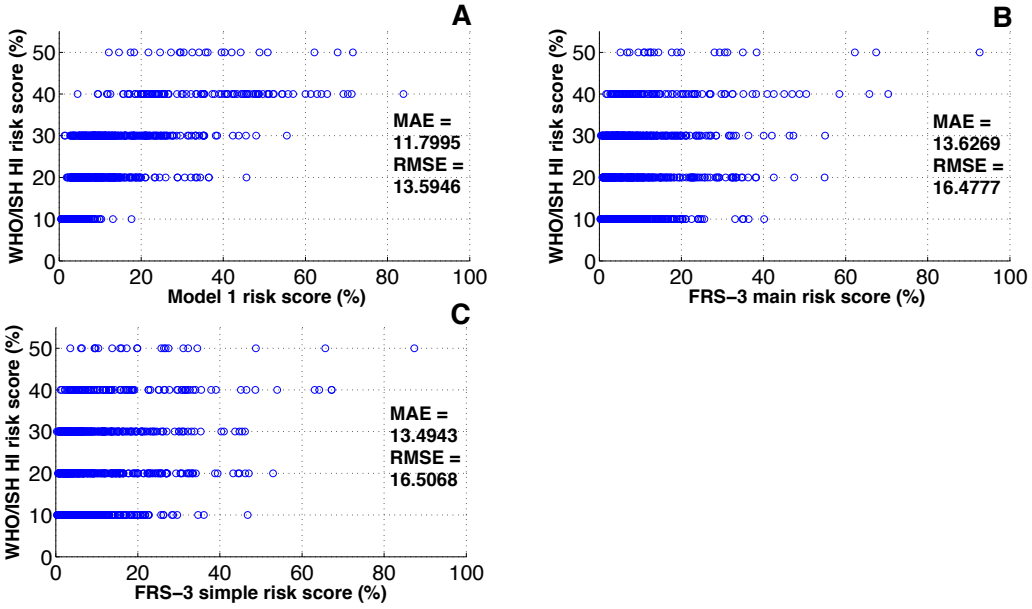


Figure 6.5 Performance of Model 1 (plot A), $FRS-3_{main}$ (plot B), and $FRS-3_{simple}$ (plot C) with the WHO/ISH high information charts on data from rural India. Both versions of the $FRS-3$ have higher MAE and RMSE, and do not show good calibration with the WHO/ISH HI risk prediction charts as compared to Model 1.

6.4 Discussion

Features were selected that were highly predictive of CVD risk. Comparing the 8 features used for building Model 1 with Table 4.1 in Chapter 4 that summarised features used in major CVD risk prediction algorithms in the literature, it can be seen that waist-to-hip ratio, CV medication - calcium channel blockers, and oral hypoglycaemics do not figure as features in any major algorithm. However, their relevance to CVD risk has been mentioned in the literature. For instance, the waist-to-hip ratio has been shown to be significantly associated with incident CVD risk [176]. Calcium channel blockers are also known to be an effective

treatment option for hypertension and CVD risk as they can be administered as monotherapy or in combination therapy and is well-tolerated [177]. Also, there is evidence that oral hypoglycaemics contributes to CVD mortality and/or morbidity [178]. The LR classifier of Model 1 has higher AUC and log-likelihood scores than the two versions of *FRS-3*. However, the performance is not substantially better than either of these two models. This may be attributed to two factors. Firstly, the usage of cross-sectional data (as opposed to utilising data over time) for prediction of CVD risk in Model 1 because of the reasons outlined in Section 6.2.2. Secondly, *FRS-3* was trained on the cross-sectional dataset that was used in this analysis which biases the comparison. However, as in the literature on CVD risk prediction algorithms described previously in Chapter 4, the AUCs obtained by Model 1 and *FRS-3* are consistent with the reported AUCs of various studies that have validated the Framingham risk score [179][180].

It was previously detailed in Chapter 4 that Cox regression was the most widely used CVD risk prediction technique. Although Cox regression can help identify the time to CVD event, for our purposes, as the prediction period (within 10 years) was fixed and cross-sectional data (which does not require censoring) was used, LR is a suitable classification technique. Also, the exact time of development of a CVD event is not as important as the probability of risk within a specified time period. One of the key difference between the two techniques is that the Cox regression specifies the instantaneous risk (or hazard) while LR specifies the proportion of patients at risk (by the odds ratio).

Another interesting point of discussion is that the contribution of age and gender alone can lead to an AUC close to 0.70, as was demonstrated with the LR classifier of Model 1. The addition of other parameters derived after a feature selection process only provides a relatively minimal improvement. This finding, however, is consistent with those reported

in article by Wald et al. [181], who claimed that CVD risk screening when performed exclusively with age offered a cost-effective method, with a performance similar to that of *FRS-3*. Their conclusion caused much skepticism and was reported as both 'radical' and 'unsurprising' [182]. Closely aligned with this finding is the research on a 'poly-pill' for CVD management. The poly-pill is a combination of statins, aspirins, blood pressure lowering drugs, and folic acid which has been manufactured in India [182]. Whilst two clinical trials have shown its effectiveness in lowering multiple risk factors such as blood pressure and LDL cholesterol [183][182], the poly-pill is a subject of active research on public health and preventive strategies for lowering the burden of CVD. Further details can be found in the article by Lonn et al. [184]. It should be mentioned that the approach based on age alone for mass drug administration, such as everyone above 55 years being prescribed a poly-pill, has come under criticism (refer to the discussion in the article by Smith et al. [182]). The mere use of age, or age and gender for CVD risk screening is a complex question with a latency between evidence and action [182].

Model 1 has lower MAE and RMSE than *FRS-3* on data from rural India when the WHO/ISH HI is used as the basis for performing a comparison. The WHO/ISH risk prediction charts are tailored for CVD risk prediction in India as we have previously described. The 10-year probabilities of CVD risk from Model 1 could be 're-calibrated' to rural India. The task of recalibrating a risk model for India is motivated by the fact that there is a lack of cohort data with recorded outcomes on CVD in India. A first approach to improve generalisation would be to modify the prognostic model for the target population [185]. Techniques for recalibration of a prediction model particularly to applications in primary care can be categorised into three [186]: recalibration by region, recalibration by ethnicity, or inclusion of additional features. As mentioned in Section 4.2.4 of Chapter 4, recalibration by ethnicity has been performed most notably on the Framingham and QRISK scores for ethnic minorities

in USA and Britain. For instance, Aarabi and Jackson [187] recommend increasing the age by 10 years when calculating the CVD risk of south Asians. A prominent example of region-specific recalibration are the WHO/ISH CVD risk prediction charts. Although little is known about the methodology of development, the charts utilise data from the Global Burden of Diseases study [117] in order to adapt the charts to different regions [186].

The Framingham risk score for prediction of Coronary Heart Disease (CHD) by Wilson et al. [87] (described previously as *FRS-2*) has been adapted for different regions of the world including Spain [188] and India [189]. The recalibration techniques rely on incorporating an adjustment of two factors - the prevalence (the number of individuals who have existing CHD in a population) and incidence (the number of individuals who will develop CHD within a given period). *FRS-2*, which was based on Cox regression, was adjusted for incidence through substitution of the baseline hazard and for prevalence through subtraction of the mean values of the model covariates in the exponentiation factor.

Although recalibrated models for CHD exist, they are limited for CVD. In the CHD recalibration process in India, the authors obtained data from mortality records to record incidence rates. A similar data collection approach is cumbersome for CVD because it encompasses a multitude of diseases including cerebrovascular diseases, peripheral vascular diseases etc. and acquisition of accurate data for causes less well understood in rural regions is challenging. Recalibration by additional risk factors is based on the premise that inclusion of extra features would explain population differences. The premise of this assumption has not been justified by concrete evidence but studies have shown different risk factors to be more predictive in certain populations [190].

Given a matrix of M features and N patients, and a vector of 10-year CVD outcomes y, we first consider the task of training a parametric model that generates β coefficients. In the Framingham dataset, it is straightforward to build a prediction model since we have no unknown entities unlike data from rural India (where outcomes are unknown). The lack of published data on incidence rates is limited. Moreover, there is likely to be an enormous variation across regions in India as exemplified by Chow et al. [189] who compared national statistics with regional statistics for CHD incidence and found a vast difference. Yusuf et al. [191] reported an income-adjusted rate of a major CVD event across four low-income countries which included India, Pakistan, Bangladesh, and Zimbabwe. If we assume that the strength of association between the features are the same across both populations, the external piece of information on incidence rate can be used to perform a first-order recalibration of the Framingham model and adapt it for India.

The 10-year probability of CVD risk from the LR classifier of Model 1 for a patient i is given by

$$\hat{p}_i = logit^{-1}(\beta_{frs} + \beta_1 X_1 + ... + \beta_m X_m) \tag{6.16}$$

To recalibrate this equation to rural India, an unknown constant β_{in} may be added to obtain

$$\hat{p}_i = logit^{-1}(\beta_{in} + \beta_1 X_1 + ... + \beta_m X_m) \tag{6.17}$$

Although individual patient outcomes on data from rural India are unknown, the overall incidence can be utilised when we consider the sum of all predicted probabilities of the

patients. For N patients, we define

$$\sum_{i=1}^{N} \hat{p} = \text{Incidence rate } \times N \tag{6.18}$$

$$\text{Incidence rate } \times N = \sum_{i=1}^{N} logit^{-1}(\beta_{in} + \beta_1 X_1 + ... + \beta_m X_m) \tag{6.19}$$

A large baseline dataset (*Dataset-3*) with 62149 participants was used to perform the recalibration but the following features were excluded from the logistic regression equation because of lack of information in the dataset: HDL, waist-hip ratio, and mother-died due to heart disease. The value of the recalibrated intercept, β_{in} was obtained to be -5.9 while the intercept from the LR equation of Model 1 that is tuned to the data from Exam 6 of the Framingham study, β_{frs}, is -6.6. The effect of these intercept values on the number of people who will be experience a CVD event in rural India is shown in Figure 6.6. The values of coefficient β_{in} and the intercept for Model 1, β_{frs} are marked on the sigmoidal curve in Figure 6.6. If the LR equation of Model 1 is used with $\beta_{in} = -6.6$, then 4530 people out of 62149 will experience a CVD event in 10-years. However, if $\beta_{in} = -5.9$ of the recalibrated equation is used, 7808 people will develop CVD in 10 years. This recalibration is, of course, sensitive to the external information on incidence rate used. The fact that the incidence rate has been generalised to four low-income countries may not be accurate for rural India.

The recalibration accounts for an offset for predicted probabilities but also relies heavily on the strong assumption of identical β coefficients. For example, the dataset descriptions in Chapter 4 reveal substantially higher smoking rates for males yet lower for females in rural India when compared to the Framingham dataset. However, given the limited availability of recorded CVD events in rural India, this approach may be preferential to an uncalibrated model.

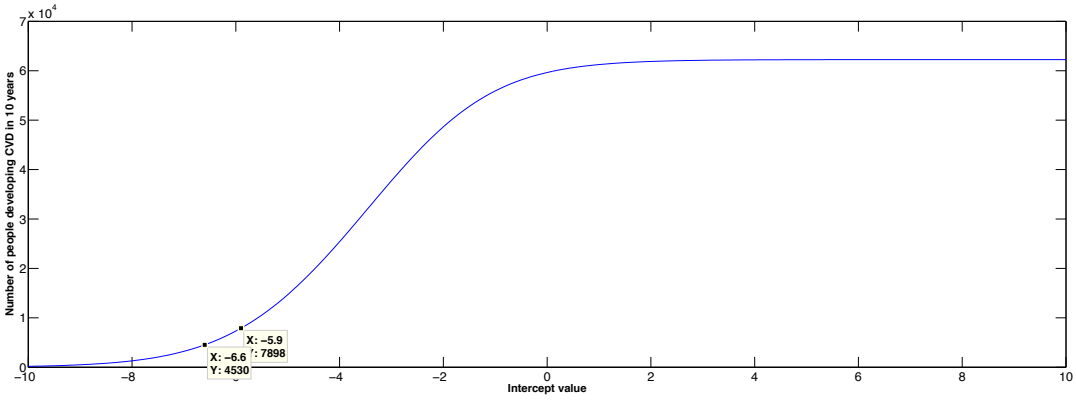


Figure 6.6 Illustrating the variation of the recalibrated intercept value and the subsequent effect on the number of people who will develop CVD in 10 years in India. The two points marked show that if an intercept of -6.6 was considered, 4530 people will develop CVD in 10 years in rural India while if an intercept of -5.9 was considered, 7808 people will develop CVD in 10 years.

6.5 Conclusion

Eight highly predictive features were identified from Exam cycle 6 of the Framingham Offspring cohort, which were similar to *the type of inputs* required by our mobile-based CDSS, *SMARThealth*. The prediction model subsequently built using the 8 features, Model 1, performed comparably to the two versions of *FRS-3* as exemplified by higher or equal AUC (0.748±0.03 for Model 1; 0.720±0.01 for FRS-3 simple; 0.729±0.01 for FRS-3 main), higher log-likelihood score (-239.99±9.59 for Model 1; -251.70±4.23 for FRS-3 simple; -247.04±4.00 for FRS-3 main), and lower Brier score (0.093±0.004 for Model 1; 0.097±0.001 for FRS-3 simple; 0.095±0.001 for FRS-3 main). The strong contribution of age and gender alone was found to lead to an AUC close to 0.70, as was demonstrated with the LR classifier of Model 1. Model 1 had a lower error rate (lower MAE-11.79 and RMSE-13,59) and better calibration when compared to the two versions of *FRS-3* when applied on data from rural India. This analysis was performed with the WHO/ISH high information charts as the benchmark algorithm. A first-order recalibration for Model 1 was presented as an approach to provide a tailored risk prediction equation to rural India.

Chapter 7

Unsupervised CVD risk model for rural India

7.1 Overview

In the previous chapter, an approach to a first-order recalibration of a CVD risk equation for rural India was presented. Although the recalibration adjusted for the difference in incidence rates (or the number of people who will develop CVD within a specified time period) between the two populations we considered, namely those in the datasets from rural India and Framingham, USA, it did not account for differences in the prevalence of key CVD risk factors between the populations. Given the lack of cohort data with recorded CVD outcomes in rural India, an unsupervised clustering approach may be more suitable to distinguish patients at high CVD risk from those with intermediate/low CVD risk. The aim of this chapter is to investigate whether the key features such as age, blood pressure, smoking status etc. cluster according to CVD risk. A subsidiary aim is to find out whether the clustering is consistent across both the rural Indian and Framingham populations.

7.2 Data description

Data was chosen from the Framingham study and the SMARThealth-baseline study, both of which were described in detail in Section 4.3.1 of Chapter 4 as *Dataset-2* and *Dataset-3*, respectively. Across both datasets, subjects with CVD at baseline (previous event) were excluded from analysis because according to clinical judgement, those who had experienced an adverse CVD event remain classified as high risk for life. In the Framingham study, the 3040 subjects from exam cycle 6 who were free of baseline CVD were selected (out of the total 3454 subjects). We shall refer to the set of 3040 patients as the *Framingham* dataset. The SMARThealth-baseline study registered 62,194 participants of whom 59027 had no previous history of CVD. We shall refer to the data from the SMARThealth-baseline study as the *rural Indian* dataset. CVD events at the end of 10 years after baseline data collection were available only for the Framingham study. The number of adverse CVD events in the *Framingham* dataset across a 10-year period is shown in Table 7.1. The adverse CVD events included the following conditions: coronary heart disease, myocardial infarction, angina pectoris, coronary insufficiency (unstable angina), cerebrovascular accident, and transient ischaemic attack.

Table 7.1 Rate of CVD events across 10 years in Exam 6 of the Framingham offspring dataset ($N_f = 3040$). The event rate is includes the events up to and including the stated year.

Year	CVD events, n (%)
1	34 (1.1)
2	72 (2.4)
3	106 (3.5)
4	125 (4.1)
5	172 (5.7)
6	206 (6.8)
7	246 (8.1)
8	281 (9.2)
9	314 (10.3)
10	345 (11.3)

The previous chapter discussed selection of highly predictive features for CVD risk. However, in the *rural Indian* dataset, features such as wasit-hip ratios could not be measured. Therefore seven features that were key CVD risk factors and measured consistently across both the *Framingham* and *rural Indian* datasets were chosen for analysis. The feature set included age, gender, systolic blood pressure (SBP), diastolic blood pressure (DBP), antihypertensive therapy, diabetes status, and smoking status. The reason for the inclusion of both SBP and DBP was exclusively because of the physiological importance of both BP components. According to Strandberg and Pitkala [192], the inclusion of both SBP and DBP improves CVD risk prediction although they recommend that SBP should be the exclusive target of antihypertensive therapy. Franklin et al. [193] also found that prediction of cardiovascular disease risk improved when both DBP and SBP were included even though SBP was a better predictor of CVD risk than DBP. Furthermore, the JNC 7 report (Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure) [64] advised inclusion of both SBP and DBP in the definition as well as the management of hypertension. Schillaci et al. [194] in their article on whether DBP should be discarded whilst assessing CVD risk, recommended the inclusion of DBP. The rationale was because of conditions such as isolated diastolic hypertension (SBP<160 and DBP$\geq$90 mmHg) which is commonly seen in younger age groups [195] or isolated diastolic hypotension (SBP$\geq$100 and DBP<60 mmHg), a condition which has been attributed to antihypertensive therapy and is an independent risk factor for heart failure. The dataset was normalised and scaled using the z-score (which involves subtraction of the mean value μ and division by the standard deviation σ across every feature).

7.3 Methods

The archetypal application of unsupervised learning is the problem of clustering data. For a given dataset, there are primarily two unknown entities of interest in an unsupervised approach: the number of clusters K, and the membership of every data point x_i to a cluster.

In the literature, clustering is a vast area of research. Theodoridis and Koutroumbas [196] describe clustering techniques alone in around 240 pages. There are two types of clustering based on the resulting output: flat clustering or partitional clustering where data is partitioned into disjoint sets, and hierarchical clustering, where a sequence of nested partitions are created [141]. Hierarchical clustering is computationally slower than flat clustering [141]. However, hierarchical clustering does not require the number of clusters to be pre-determined in contrast to flat clustering algorithms such as the k-means algorithm which have been shown to be sensitive to initial conditions as well as requiring a formal method of selecting K [141]. There are two main approaches to hierarchical clustering: a top-down or *divisive clustering* approach that starts from one big cluster that is progressively fragmented, and a bottom-up or *agglomerative clustering* approach. Agglomerative clustering has been more frequently used in practice than the divisive approach [141].

7.3.1 Agglomerative clustering

Agglomerative clustering is an iterative process that starts with N singleton clusters. At each step of the iteration, the two most 'similar' clusters are merged and this process is continued until there is one big single cluster containing all data samples. The 'similarity' is the measure of distance between two data points or clusters and is stored as a matrix S. Often, a monotonically decreasing function is applied to S to obtain a dissimilarity matrix D where $d_{ii} = 0$ and $d_{ij} \geq 0$ for two data samples x_i and x_j. Similarity and dissimilarity essentially

convey the same notion of distance but the interpretation for the former is reciprocal to distance, i.e. the higher the similarity, the smaller the distance. There are several distinct metrics for computing similarity or dissimilarity. A full review of these metrics is presented in Gan et al. [197]. Depending on the type of data (categorical, ordinal, continuous, mixed), different metrics are available. For instance, the generalised similarity coefficient [198] and the generalised Minkowski distance [199] have been used in the literature for mixed data types. In this chapter however, a rudimentary approach is followed initially with the widely used squared Euclidean distance metric. This is also a pre-requisite for our choice of 'linkage function', which is described below. Given two data points x_i and y_i with $i = \{1...m\}$ dimensions, the Euclidean distance is defined as $\sum_{i=1}^{m} \sqrt[2]{(x_i - y_i)^2}$, and the squared Euclidean is defined as $\sum_{i=1}^{m} (x_i - y_i)^2$.

The distance between two clusters is measured by a 'linkage' function. A convenient approach to computing distance is by either taking the nearest points between the two clusters ('the single linkage method'), or the farthest point in each cluster ('the complete linkage' method) [141]. The average distance between all pairs of points could also be considered and this method is referred to as the average linkage method [141]. Another procedure to compute cluster distance is through Ward's minimum variance method [200], which imposes an objective function that is used as a criterion to select a pair of clusters for merging at each step. This implementation considers the clustering problem as an analysis of variance. The algorithm is described in Figure 7.1. The Error Sum of Squares (ESS) captures the mean within clusters for each feature while the Total Sum of Squares (TSS) considers the grand mean of individual features.

To visualise the process of merging data points or clusters, a plot known as a dendrogram is constructed. It is quintessentially a binary tree where initial groups are found at the bottom

procedure AGGLOMERATIVE CLUSTERING- WARD'S MINIMUM VARIANCE METHOD
  Given data X of n patients and m features;
  initialize clusters as singletons: **for** $i \leftarrow 1$ **do** $C_i \leftarrow \{i\}$;
  initialize the cluster set available for merging $S \leftarrow \{C_1, ..., C_n\}$;
  repeat
    Compute intra-cluster variance, the Error Sum of Squares, $ESS = \sum_{i=1}^{S}\sum_{j=1}^{n}\sum_{k=1}^{m}\left|X_{ijk} - \bar{x}_{ik}\right|^2 \forall$ clusters $i = \{1, ..., S\}$, patients $j = \{1, ..., n\}$, and features $k = \{1, ..., m\}$; $\bar{x}_{ik}$ is the mean of the k^{th} feature vector in the i^{th} cluster.
    Compute inter-cluster variance, the Total Sum of Squares, $TSS = \sum_{i=1}^{S}\sum_{j=1}^{n}\sum_{k=1}^{m}\left|X_{ijk} - \bar{x}_{k}\right|^2 \forall$ clusters $i = \{1, ..., S\}$, patients $j = \{1, ..., n\}$, and features $k = \{1, ..., m\}$; $\bar{x}_{k}$ is the mean of the k^{th} feature vector.
    Define r relating ESS and TSS as $r^2 = \frac{TSS-ESS}{TSS}$;
    Select 2 clusters to merge based on dissimilarity $(C_a, C_b) \leftarrow max\ r^2$;
    Merge clusters $C_l \leftarrow C_a \cup C_b$;
    Annotate (C_a, C_b) as unavailable $S \leftarrow S \backslash \{C_a, C_b\}$;
    If C_l does not include n patients, then $S \leftarrow S \cup C_l$;
  until $C_l = \{1, ...n\}$;
end procedure

Figure 7.1 Ward's minimum variance method for hierarchical clustering

('leaves') with the root containing all data samples. As groups are merged, they progressively join to form trees. The height of the branches represent the extent of dissimilarity between groups. The desired number of clusters K does not need to pre-determined in hierarchical clustering and inspection of the dendrogram aids visual intuition with regard to a suitable K. However, visible 'gaps' may be difficult to detect in many circumstances and therefore choosing the 'correct' K requires a more quantitative basis for justification.

7.3.2 Estimation of optimal K

There are several techniques for estimating the optimal value of K. A full review can be found in the article by Gordon et al. [201], who categorised methods of estimating the optimal K, or $\hat{K}$ into two: global and local methods. Global methods define a metric utilising the entire dataset and optimise it on the number of clusters. Local methods consider a pair of clusters

at a time to determine if they need to be grouped. A key problem with the global method is that it can treat the entire dataset as one cluster and offers no guidance on whether data should be clustered. However, this can be overcome based on prior assumptions about the existence of clusters present in the dataset.

One such global method was developed by Davies and Bouldin [202] who proposed a metric that optimises the *intra-cluster* scatter S or how closely grouped the clusters are, and the *inter-cluster* variation M or how much the clusters are spread apart from each other. These are given by

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} \left| X_j - A_i \right|^q \right\}^{\frac{1}{q}} \tag{7.1}$$

$$M_{ab} = \left\{ \sum_{k=1}^{N} \left| A_{ka} - A_{kb} \right|^p \right\}^{\frac{1}{p}} \tag{7.2}$$

where S_i and M_{ab} are the intra-cluster scatter and inter-cluster variation respectively for the i^{th} cluster, j^{th} patient, k^{th} feature, and for clusters a and b. X_j represents the data comprising all features for the j^{th} patient. A_{ka} is the centroid of cluster a for the k^{th} feature, A_i is the centroid of cluster i for all features, and T_i is the size of cluster i. M_{ab} is the Minkowski distance between the centroids of clusters a and b. When p=1, M_{ab} equates to the city-block distance between the centroids of clusters a and b. When p=2, M_{ab} reduces to a Euclidean distance measure between centroids of clusters a and b. The q^{th} root of the q^{th} moment of all points in cluster i about their mean is given by S_i. When the value of q is 1, S_i transforms to the Euclidean distance between X_j and A_{ki}. When q takes the value of 2, S_i becomes the standard deviation of the distance of all patients in a cluster to the respective cluster's centroid. Combining the equations described in 7.1, we obtain

$$R_{ij} = \frac{S_i + S_j}{M_{i,j}} \tag{7.3}$$

$$\bar{R} = \sum_{i=1;i\neq j}^{N} max\ R_{ij} \tag{7.4}$$

where R_{ij} represents a ratio of scatter *within* clusters i and j to scatter *between* clusters i and j. Lower values of $\bar{R}$ indicate clusters that are compact with greater separation between the centroids of clusters. Therefore the number of clusters for which $\bar{R}$ is minimum reveals the optimal number of clusters.

Calinski and Harabasz [203] also derived a ratio between the inter-cluster variation and intra-cluster scatter as a metric to estimate $\hat{K}$.

$$CH_K = \frac{M(K)/(K-1)}{S(K)/(n-K)} \tag{7.5}$$

For a specified number of clusters (e.g. 20), $\hat{K} = argmaxCH(K)$ over $K \in 2, ...K_{max}$. CH is undefined for $K = 1$ as may be observed from Equation 7.5.

The Gap Criterion, proposed by Tibshirani et al. [204], utilises the information provided by the decrease in the intra-cluster scatter over several possible values of K. It determines the cluster most resilient to random perturbations [205]. The intra-cluster scatter, $S(K)$, is compared to $S_{unif}(K)$ which is the intra-cluster scatter if the samples were distributed uniformly such that

$$Gap(K) = \log\ S_{unif}(K) - \log\ S(K) \tag{7.6}$$

The term $logS_{unif}(K)$ is determined through Monte Carlo simulations. An average value is computed with 20 simulated uniform datasets [204], and a standard error of $logS_{unif}(K)$ is derived such that

$$\hat{K} = \min \left\{ Gap(K) \geq Gap(K+1) - s(K+1) \right\} \tag{7.7}$$

where $K \in \{1, \ldots K_{max}\}$.

The methodology of our analysis was programmed and implemented using MATLAB [206], and is summarised in Figure 7.2. Outcomes were known only for the *Framingham* dataset. Therefore the risk bands from WHO/ISH CVD risk prediction charts were applied to each cluster formed for the Indian dataset, acting as a proxy measure of 10-year CVD events.

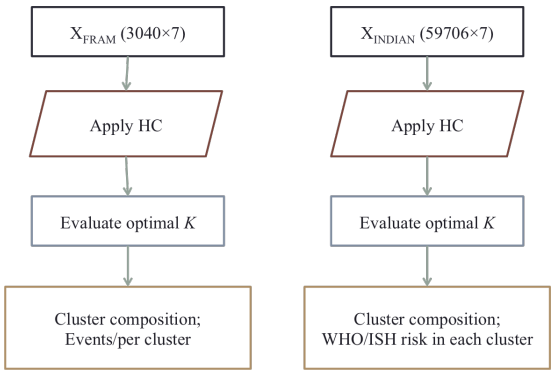


Figure 7.2 Methodology for clustering patients with the Framingham and Indian datasets. The *Framingham* dataset consists of 3040 patients with 7 features and the *rural Indian* dataset consists of 59706 participants across the same 7 features. HC stands for hierarchical clustering. The optimal K was evaluated using three criteria, namely Davies and Bouldin, Calinski and Harabasz and Gap.

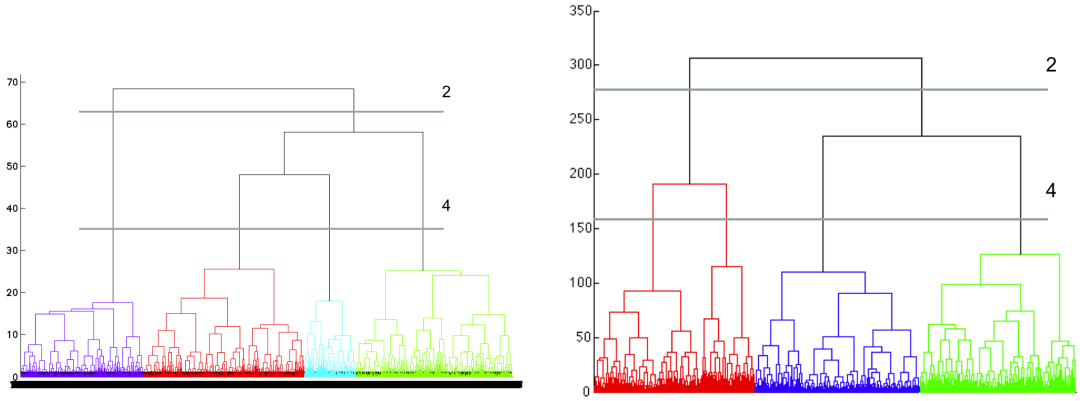
7.4 Results and discussion

7.4.1 Optimal number of clusters

Upon application of hierarchical clustering to both sets of data, we obtain two dendrograms as illustrated in Figure 7.3. A suitable K can be inspected from the height of the branches that convey the distance between the clusters. Through visual inspection, it appears suitable to cut the dendrograms at K = 2 or 4. A more objective approach however, would be to determine K through cluster evaluation metrics. Figure 7.4 shows the results of applying three metrics namely, the Davies-Bouldin (DB), Gap Criterion (GC), and Calinksi-Harabaz (CH) on the *Framingham* and *rural Indian* datasets for different values of $K \in \{1,...K_{max}\}$. The GC and CH criteria indicate the optimum K through the highest value while the DB criterion optimises the minimum value. It may be observed that for the *Framingham* dataset, $\hat{K}_{fCH} = 2$, $\hat{K}_{fGap} = 4$, and $\hat{K}_{fDB} = 4$. For the *rural Indian* dataset, $\hat{K}_{inCH} = 3$, $\hat{K}_{inGap} = 2$, and $\hat{K}_{inDB} = 4$.

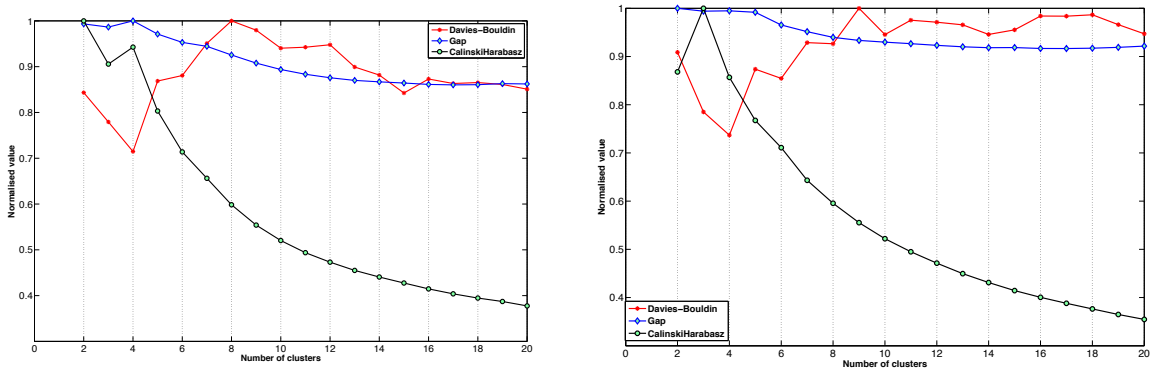
The lower branches and roots of the dendrograms are coloured by the default MATLAB implementation which uses an arbitrary threshold that is 70% of the maximum linkage computed between clusters. However, we observe that the part in red colour in the *rural Indian* dataset actually comprises of two groups even though it has been coloured as one. It may also be observed that $\hat{K}$ =2, 3, and 4 are broadly comparable when based on the Gap Criterion.

Fundamentally, CVD risk stratification can be formulated as a dichotomous problem because we are interested in separating those patients at high risk of CVD from those with a lower CVD risk to provide two different treatment approaches. D'Agostino et al. [88] who developed the Framingham risk score, chose 3 cut-offs for categorising risk - 0% to 6%, 6% to 20%, and >20%, while the WHO/ISH risk prediction charts for India suggest 5 bands of risk



(a) *Framingham* dataset (N_f=3040) (b) *rural Indian* dataset (N_{in}=59706)

Figure 7.3 Dendrograms representing a hierarchical clustering tree for the *Framingham* and *rural Indian* datasets. The two lines that cut the dendrogram help to visualise an appropriate choice of K. Lower roots and branches of the dendrograms are coloured by thresholding at 70% of the maximum linkage computed between clusters. It can be observed that the red coloured part in the *rural Indian* dataset actually comprises of two groups.



(a) Optimal K for the *Framingham* dataset (b) Optimal K for the *rural Indian* dataset

Figure 7.4 Selection of the optimal number of clusters K through the Davies-Bouldin, Gap, and Calinski-Harabasz criteria. This vertical axis of figure represents a normalised index. To determine optimal K, both the Gap and CH criteria use the highest indexed value while the Davies-Bouldin criterion utilises the minimum value. For the *Framingham* dataset, $\hat{K}_{fCH} = 2$, $\hat{K}_{fGap} = 4$, and $\hat{K}_{fDB} = 4$. For the *rural Indian* dataset, $\hat{K}_{inCH} = 3$, $\hat{K}_{inGap} = 2$, and $\hat{K}_{inDB} = 4$.

[1]. From Figure 7.4, we observe that for $\hat{K} = \{2,4\}$, all 3 evaluation metrics show comparable values. Table 7.2 describes the average composition of CVD events over $t = \{1,2,..10\}$ years to identify the optimal value of K for the *Framingham* dataset which has recorded outcomes. The average event rate can be defined as the sum of CVD events (previously defined in 7.2) over $t = \{1,2,..10\}$ years for each cluster, given by $events_i = \frac{1}{T}\sum_{t=1}^{10}\sum_{j=1}^{N_i} e_{ijt}$ where $T = \sum_{t=1}^{10}\sum_{j=1}^{N} e_{jt}$, and $e \in [0,1]$ represents a CVD event or non-event for the i^{th} cluster, j^{th} patient and N is the number of patients. Values for $\hat{K} = \{2,3,4,5\}$ clusters are considered.

Table 7.2 Average CVD event rate for the *Framingham* dataset for $\hat{K} = \{2,3,4,5\}$. It is observed that when $\hat{K} = 2$, the second cluster has an overwhelming majority (77%) of the CVD events.

Cluster/ Average Event Rate	$\hat{K} = 2$ in % (N)	$\hat{K} = 3$ in % (N)	$\hat{K} = 4$ in % (N)	$\hat{K} = 5$ in % (N)
1	22.14±2.82 (1523)	22.14±2.82 (1523)	10.07±1.30 (779)	14.07±2.65 (467)
2	77.86± 2.82 (1517)	31.81±4.79 (741)	18.21±2.13 (844)	17.77±3.23 (274)
3		46.03±7.27 (776)	22.17±3.82 (441)	2.79±0.73 (514)
4			49.59±6.22 (976)	19.35±2.91 (1009)
5				46.03±7.27 (776)

From Table 7.2, it is observed that when $\hat{K} = 2$, the second cluster has the overwhelming majority (77%) of the CVD events. In the case of $\hat{K} = 3$, the sole difference is that the 77% events found in cluster 2 of $\hat{K} = 2$, seems to be bifurcated as clusters 2 and 3 in $\hat{K} = 3$. In

the case of $\hat{K} = 4$, the emergence of a 'low-risk' cluster, or a cluster with low event rates of 10% is observed. For $\hat{K} = 5$, the CVD event rate for clusters 1,2, and 4 is comparable to the 10-year event rate in *Framingham* dataset of 11.3%, indicating no clear separation of events. Also, $\hat{K} = 5$ is not favoured by any of the three criteria for selection of the optimal value of K (see Figure 7.4). Therefore for further analysis, values of $\hat{K} = \{2, 4\}$ are considered.

7.4.2 Cluster composition

Tables 7.3 and 7.4 describe the mean levels of risk factors of the datasets used as input for clustering. The mean and standard deviation of the risk factors are reported for $\hat{K} = 2$ and $\hat{K} = 4$ and the patterns of their spread across different clusters may be observed. For $\hat{K} = 2$, the resultant clusters have markedly different levels of age, SBP, DBP, antihypertensive therapy, and diabetes. Gender and smoking status appear to be closer to the mean levels of the overall data. For $\hat{K} = 4$, the clusters appear to be a more refined version of those clusters that were observed when $\hat{K} = 2$. For instance, from Table 7.3, clusters 1 and 2 for $\hat{K} = 4$ could be interpreted as an expansion of cluster 1 for $\hat{K} = 2$. Similarly, clusters 3 and 4 for $\hat{K} = 4$ appear to be an expanded version of cluster 2 for $\hat{K} = 2$. This is exemplified by the BP values in Table 7.3 as BP for N_{f21} and N_{f22} are approximately an average of the values for N_{f41} and N_{f42} and N_{f43} and N_{f44}, respectively. However, the distribution of the averaged percentage of smokers in N_{f21} and N_{f22} are opposite to that of the pattern of distribution of BP. For example, N_{f41}, which has younger participants, more females, lower levels of BP, antihypertensive therapy, and diabetes, than cluster 2, records higher smoking rates.

There are, however, some key differences across the populations. For instance, smoking status is skewed for the *rural Indian* dataset with a prevalence of over 40% amongst males and 5% amongst females, as described in Chapters 4 and 3. There is however, coherence and

consistency in the cluster formations as exemplified by the increase in mean risk factor levels as one goes from clusters 1 to 4. The event rates in the different clusters are presented in the next section.

Table 7.3 Mean risk factors across clusters in *Framingham* dataset

$\hat{K}$	Feature	Age (years) Mean±std	% female	SBP (mmHg) Mean±std	DBP (mmHg) Mean±std	Antihypertensive therapy (%)	Diabetes (%)	Smoking (%)
	Entire dataset (N_f=3040)	57.85 ± 9.51	55.33	127.71 ± 18.52	75.65 ± 9.36	24.11	6.58	15.72
2	Cluster 1 (N_{f21}=1523)	52.17 ± 7.19	58.9	115.60 ± 11.08	72.19 ± 7.70	8.96	2.1	17.4
	Cluster 2 (N_{f22}=1517)	63.55 ± 8.03	51.75	139.88 ± 16.41	79.13 ± 9.59	39.32	11.07	14.4
4	Cluster 1 (N_{f41}=779)	52.11 ± 7.49	59.57	109.25 ± 9.10	66.41 ± 5.49	1.74	0.43	33.04
	Cluster 2 (N_{f42}=844)	51.48 ± 5.59	56.49	123.29 ± 10.55	79.04 ± 6.19	20.26	5.56	12.98
	Cluster 3 (N_{f43}=441)	66.03 ± 5.95	61.7	127.25 ± 12.23	71.09 ± 6.43	26.38	8.79	3.68
	Cluster 4 (N_{f44}=976)	64.55 ± 7.97	44.23	150.84 ± 16.01	83.83 ± 8.62	45.07	11.95	17.61

Table 7.4 Mean risk factors across clusters in *rural Indian* dataset

$\hat{K}$	Feature	Age (years) Mean±std	% female	SBP (mmHg) Mean±std	DBP (mmHg) Mean±std	Antihypertensive therapy (%)	Diabetes (%)	Smoking (%)
	Entire dataset (N_{in}=59706)	53.81 ± 10.95	53.68	125.64 ± 22.05	79.53 ± 11.67	29.00	17.93	25.38
2	Cluster 1 (N_{in21}=39835)	54.07 ± 11.15	48.99	115.67 ± 13.83	74.54 ± 8.95	14.91	14.60	24.31
	Cluster 2 (N_{in22}=19871)	63.08 ± 10.54	58.9	145.62 ± 21.87	89.55 ± 9.91	25.05	24.61	27.53
4	Cluster 1 (N_{in41}=20530)	66.03 ± 5.95	47.91	111.13 ± 10.70	74.86 ± 7.99	00.31	12.85	21.42
	Cluster 2 (N_{in42}=13225)	48.04 ± 9.18	63.67	135.63 ± 14.89	89.96 ± 7.33	10.22	27.74	29.66
	Cluster 3 (N_{in43}=19305)	64.55 ± 7.97	50.14	120.49 ± 15.09	74.20 ± 9.87	30.44	16.47	27.38
	Cluster 4 (N_{in44}=6646)	63.72 ± 9.63	61.90	165.51 ± 19.88	88.75 ± 13.63	54.57	18.40	23.28

7.4.3 Events per cluster

Figure 7.5 demonstrates that event rate across clusters 1, 2, and 3 are approximately constant. Cluster 4 has the highest composition of CVD events, which is as high as 65% of the cluster in year 1 to 46% of the cluster in year 10. However, the percentage of events in year 1 is

low and hence the event rate in cluster 4 is exceptionally large for year 1. The percentage of events is with respect to all events that occurred during that year. It is clear that cluster 4 is 'high risk' (HR) while cluster 1 is 'low risk' (LR). Clusters 3 and 2 can be described to be 'upper intermediate risk' (UIR) and 'lower intermediate risk' (LIR). This is further elucidated by Table 7.5 which describes event rates not as a percentage of the overall rate, but as a percentage of the total number of patients. For instance, we can observe that the 11.3% event rate in 10 years is distributed as 1.2% for LR cluster, 2.4% for LIR cluster, 2.6% for UIR cluster, and 5.3% for HR cluster.

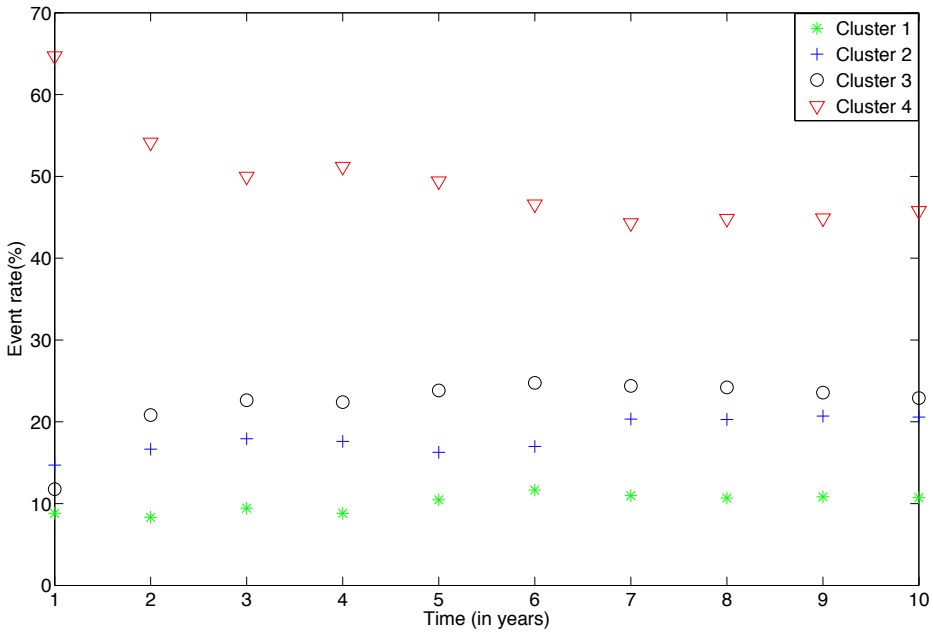


Figure 7.5 Event rate per cluster across $t = \{1, ..., 10\}$ period for the *Framingham* dataset. It is expressed as a percentage of the overall event rate for the t^{th} year. For example, cluster 4 approximately captures 50% of all CVD events across the 10 year period. It may be observed that the clusters are constant and consistent in their composition of event rates.

Table 7.6 describes the stratification of data in 5 WHO/ISH CVD risk bands. It is clear that cluster 1 has far fewer events amongst all clusters and hence is LR. Cluster 4 which has

Table 7.5 CVD events per cluster in *Framingham* dataset. † denotes the % to be relative to N_f

CVD in t^{th} year	Events in *Framingham* dataset (N_f=3040)	Cluster 1 events (N_{f41}=779) n (%†)	Cluster 2 events (N_{f42}=844) n (%†)	Cluster 3 events (N_{f43}=441) n (%†)	Cluster 4 events (N_{f44}=976) n (%†)
1	34 (1.1)	3 (0.1)	5 (0.2)	4 (0.1)	22 (0.7)
2	72 (2.4)	6 (0.2)	12 (0.4)	15 (0.4)	39 (1.3)
3	106 (3.5)	10 (0.3)	19 (0.6)	24 (0.8)	53 (1.8)
4	125 (4.1)	11 (0.4)	22 (0.7)	28 (0.9)	64 (2.1)
5	172 (5.7)	18 (0.6)	28 (0.9)	41 (1.4)	85 (2.8)
6	206 (6.8)	24 (0.8)	35 (1.2)	51 (1.7)	96 (3.2)
7	246 (8.1)	27 (0.9)	50 (1.7)	60 (2.0)	109 (3.6)
8	281 (9.2)	30 (1.0)	57 (1.9)	68 (2.3)	126 (4.2)
9	314 (10.3)	34 (1.1)	65 (2.2)	74 (2.5)	141 (4.7)
10	345 (11.3)	37 (1.2)	71 (2.4)	79 (2.6)	158 (5.3)

the largest number of patients with CVD risk over 30% and substantially fewer patients with CVD risk less than 10% is the HR cluster. Cluster 3 has more patients across risk bands 10 to 30% and is the UIR cluster. Cluster 2 has a large number of low risk patients but fewer dispersed across risk bands 10-30%, and is suitably the LIR cluster. It may also be observed that the size of the HR cluster 4 is approximately a tenth of the entire dataset. In Chapter 5 we described the *rural Indian* national guidelines for management of CVD risk at primary care as advocated by the 2009 Indian National Programme For Prevention and Control of Cancer, Diabetes, Cardiovascular Diseases & Stroke (NPCDCS). The criteria for treating high-risk patients is based on a combination of risk bands 30-40% and greater than 40%, which implies that four bands of CVD risk in the WHO/ISH charts are suitable as well. The distribution of CVD risk across different clusters in the *rural Indian* dataset is shown in Table 7.6,

Implications of findings Identification of the 'correct' K is one of the difficult aspects of cluster analysis as there is no clear definition of a cluster [204]. In our approach, we have utilised three criteria for selecting an optimal K, which was also aided by an understanding of the number of 'risk categories' we wished to obtain. Hierarchical clustering is known to

Table 7.6 Distribution of CVD risk across different clusters in the *rural Indian* dataset. The overall dataset column shows the distribution of 59706 participants across each CVD risk band (with the % expressed in brackets). The columns for clusters 1 to 4 show the number of participants in every CVD risk band in each cluster. The percentage within the bracket is the proportion of the number of participants in a CVD risk band within a cluster to all participants within that cluster.

10-year risk WHO/ISH	Overall dataset (N_{in}=59706)	Cluster 1 (N_{in41}=20530; 34.4%)	Cluster 2 (N_{in42}=13225; 22.2%)	Cluster 3 (N_{in43}=19305; 32.3%)	Cluster 4 (N_{in44}=6646; 11.1%)
<10%	41747 (69.92%)	20935 (99.34%)	11176 (84.51%)	9697 (50.23%)	479 (7.21%)
10 to <20%	9081 (15.21%)	110 (0.54%)	1103 (8.34%)	6958 (36.04%)	910 (13.69%)
20 to <30%	4485 (7.51%)	6 (0%)	637 (4.82%)	2022 (10.47%)	1820 (27.38%)
30 to <40%	1447 (2.42%)	4 (0%)	169 (1.28%)	411 (2.13%)	863 (12.99%)
Greater than 40 %	2889 (4.84%)	1 (0%)	123 (0.93%)	202 (1.05%)	2563 (38.56%)

produce clusters even if the data has no structure at all [141]. However, cluster evaluation methods such as the Gap criterion are specialised to pick null clusters [204]. $\hat{K} = 2$ can be useful in identifying a cluster with a smaller size (1517 as opposed to overall dataset of 3040) and much higher prevalence of CVD events (77%) than the overall prevalence rate (11.3%). $\hat{K} = 4$ generates fine grained clusters that were coherent and consistent as exemplified through mean risk factor levels and event rates. Also, for $\hat{K} = 4$, the event rates observed in the 4 clusters across $t = \{1, ..., 10\}$ period indicated the clusters were consistently identifiable (i.e. as low risk, high risk, UIR, or LIR) in the *Framingham* dataset. Although we could not verify the same in the *rural Indian* dataset due to lack of recorded outcomes, the WHO/ISH risk bands indicate a similar pattern for the 10 year events.

The clustering approach presented here can be useful in terms of providing additional information in a classification or survival model for CVD. Alternatively, prediction models may be trained specifically for each cluster which is likely to result in different features being more predictive in different clusters. In India, there exists a lack of a population-specific CVD risk model except the WHO/ISH charts. In chapter 5, we explored the WHO/ISH charts which have reported neither model performance nor validation. In the absence of a calibrated risk score, it may be sufficient to identify groups of patients with varying risk. An unsupervised

approach such as the one demonstrated here, which is validated across a population with known outcomes can be useful, although to a lesser extent than if the outcomes were known. For instance, clusters trained and programmed from an electronic health record, such as the OpenMRS system we have used in *SMARThealth*, can help identification of high-risk groups. This can be combined with parameters such as location in order to derive more region-specific insights. This will be further investigated upon completion of the *SMARThealth-RCT*, which will shall outline in the next chapter.

7.5 Conclusion

Using an unsupervised clustering technique, we have identified clusters that established low CVD risk and high CVD risk (when $\hat{K} = 2$) patients as well as upper intermediate CVD risk and lower intermediate CVD risk (when $\hat{K} = 4$). The cluster compositions, and optimal number of clusters were consistent across both the *rural Indian* and *Framingham* datasets. Furthermore, the event rates across a 10 year period were were similar when validated with recorded outcomes from the *Framingham* dataset. Unsupervised clustering offers an alternative approach to identifying groups of high-risk patients in rural India.

Chapter 8

The SMARThealth Randomised Controlled Trial

8.1 Introduction

So far, we have introduced the problem of CVD, an mHealth CDSS to enhance the abilities of ASHAs and PHC physicians to perform CVD risk assessment, and investigated different approaches to improving existing risk prediction techniques. The long-term aim is to translate improvements on to the CDSS. The process of obtaining ethical approval for a large-scale randomised controlled trial to evaluate the benefits of using the CDS in rural India was initiated before the pilot study.

The design of the RCT, whose main aim is to quantify the clinical impact of a change in blood pressure following the use of the CDSS is described. Blood pressure was used as an end point for two reasons: firstly, CVD manifests itself over a longer period and would therefore require a study duration (10 years) beyond the funding available for the RCT; secondly, BP is an important predictor of CVD risk as demonstrated in Chapter 6. We present a production-ready *SMARThealth-intervention* mobile application that consists of

improvements over the *SMARThealth* pilot and baseline applications described in Chapters 2 and 3 respectively. Finally, an analysis of intermediate data from the RCT is presented.

8.2 Design of RCT

A randomized controlled trial is the gold standard for generating evidence in biomedicine. The design for this trial is in the form of a stepped wedge cluster RCT (swc-RCT). Figure 8.1 illustrates the conventional swc-RCT, where it can be observed that clusters are sequentially randomised to receive intervention at specified time points. The trial has 3 arms, each with 18 villages and 6 PHCs. It consists of four 'phases', each of which represent a 6-month interval. The trial design utilised here is referred to as a 'complete design' swc-RCT since each cluster is scheduled to be a part of both the control and intervention groups, albeit at different time points. The descriptions of the other forms of swc-RCTs such as incomplete design (with and without an implementation period), parallel RCTs can be found in the article by Hemming et al. [207].

The primary difference between a swc-RCT and a conventional RCT is that, as opposed to individual randomisation, the stepped wedged design progressively shifts a group of individuals within villages to the intervention at different time points. It is particularly suitable for ethical reasons (since no individual is left unexposed to the intervention), and logistic reasons (all ASHAs and PHC may not need to be trained at once to use the CDSS) in our trial. The primary outcome from the SMARThealth RCT is to determine whether high-risk individuals who receive the intervention could achieve adequate control of their blood pressure levels. Secondary outcomes include mean change in other CVD risk factors such as smoking status and diabetes, quality of Life, CVD events, and process outcomes

(such as referrals to a physician).

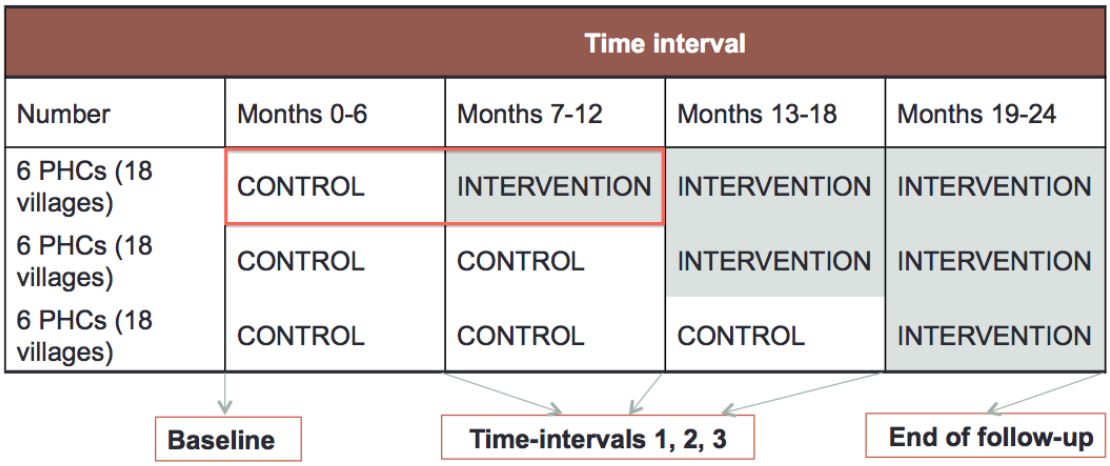


Figure 8.1 Design of stepped wedged randomised controlled trial

Participants of the RCT are people aged above 40 years who have been identified as having a high risk of CVD during the baseline study. Each cluster in the swc-RCT comprises 18 villages and 6 PHCs. The intervention is based on the mHealth CDSS and includes workforce training and field support to the ASHAs and PHC physicians. A summary of the salient features of the intervention are given below and described in detail in the next section (section 8.3).

- A back pack sized kit comprising a mobile 7-inch tablet (with the CDSS application pre-loaded in it), BP monitor, blood glucometer, weighing scales, and measuring tape.
- A shared electronic health record installed on a central server, that continuously captures and processes patient information from multiple tablet devices.
- A referral system with PHCs for patients identified to be at high risk of CVD. Barcode enabled referral cards to track patients.
- A prompt system to alert high-risk individuals for follow-up visits with ASHA / PHC physician and reminders about medication.

- Training and resource support for ASHAs and PHC physicians. This includes training for operating the CDSS, awareness workshops on management of CVD, additional remuneration for using the CDSS, and capacity enhancement of services provided by the ASHAs.

The control group receives existing health care available through the PHC without access to the mobile health component and the workforce support as described above.

Determining sample size The goal of estimating sample size (SS) is to gauge the number of participants required to detect a clinically meaningful result [208]. If the SS is too small, it is difficult to detect an important effect or discuss generalisation. However, a large SS could be a considerable strain on the time and resources needed to conduct the RCT. Four basic components are involved in the calculation of the sample size. They are summarised below:

Alpha

Alpha determines the probability of incorrectly detecting a statistically significant difference (type 1 error; false positive) between the groups of comparison.

Beta

Beta governs a false negative result, or, the probability of incorrectly rejecting a statistically significant difference between the groups of comparison.

Power

The power is a complement of beta (1-beta) and is the true positive. In other words, the power represents the probability of correctly detecting an effect present in the population based on testing a sample from that population.

Variability

The sample size is highly dependent on population variance of the outcome. This is normally estimated from a previous study or a pilot study (which in our case was the

Andhra Pradesh Rural Health Initiative [3]). An associated parameter is the Minimal Clinically Relevant Difference (MRCD), which defines the smallest difference in the outcome desired by the trial investigators. For instance, the proportion of people with optimal BP levels (which is defined by SBP<140 mmHg) was recorded as 39% in the APHRI study.

In the literature, descriptions of power calculations for swc-RCTs are limited [207]. The most important work in this area was presented by Hussey and Hughes [209], who established the theoretical basis for power calculation for a 'complete design' swc-RCT. The computation of statistical power for the SMARThealth-RCT was performed by the statisticians at the George Institute for Global Health, Sydney. As 6 PHCs and 18 villages (comprising 24,000 eligible people approximately) are included to receive the intervention every 6 months, it was possible to detect an absolute increase of 6% in the proportion of people with optimal BP levels (defined as SBP<140, DBP<90 mmHg). This could be achieved with greater than 90% power (2α=0.05) and translated to an increase from 39% to 45% of the proportion achieving optimal BP levels, as well as a mean SBP difference of 3 mmHg between the control and intervention groups. Five time-points for data collection were assumed to perform the calculations.

8.3 mHealth intervention

8.3.1 Electronic health records

The Open Medical Record System

OpenMRS is a freely available open source medical record system that was originally motivated by the need to create electronic health records in resource-poor settings. It was originally developed as the result of the collaboration between the Regenstrief Institute in Indianapolis and the Boston-based Partners-In-Health, with the aim of improving the IT

infrastructure for managing HIV/AIDS in western Kenya [210]. Over the last 10 years, OpenMRS has been used in many African and Latin American countries, and it is supported by a network of international volunteers and developers [211].

Software architecture OpenMRS is mainly built using Java and follows the Model-View-Controller (MVC) architecture. It consists of three layers: the data layer, the service layer, and the user interface layer as shown in Figure 8.2.

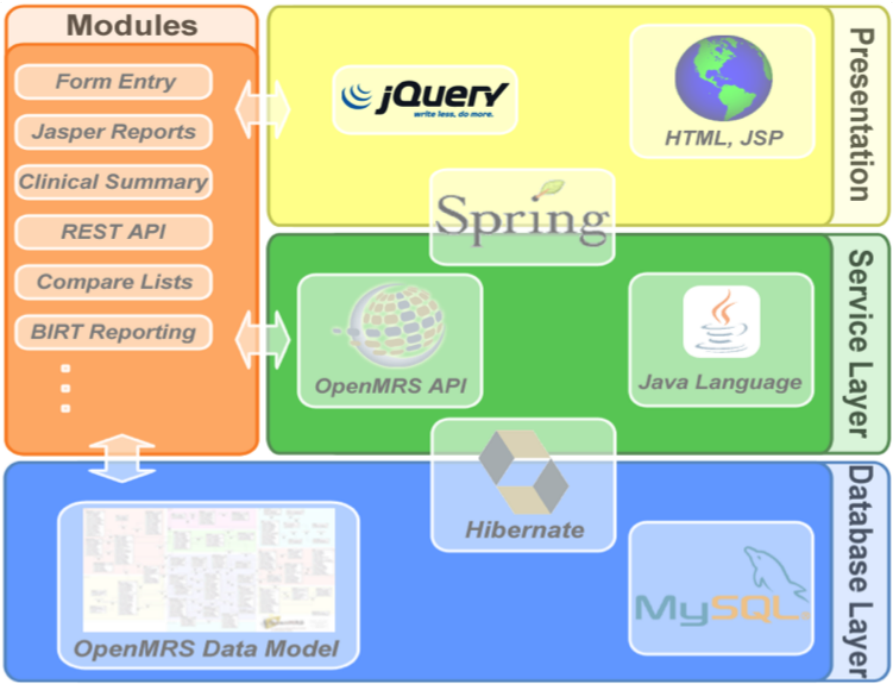


Figure 8.2 The OpenMRS architecture [212] comprises three main components, namely the data layer, the service layer, and the user interface layer. The framework is made modular so that developers may add their own modules to customise implementation. This diagram is adapted from the OpenMRS implementers guide [212]. 'Jasper Reports' is a popular open source reporting library in Java which can write to a plethora of devices (e.g. printer, screen) and formats (e.g. PDF, HTML, Comma-separated values). 'BIRT Reports' stands for Business Intelligence and Reporting Tools, and is an open source Business Intelligence and data visualisation system.

The data layer The data layer uses an object-relational mapping tool called Hibernate for abstraction of the MySQL database, as well as Liquibase to manage safe re-factoring

of the database whenever changes occur. The MySQL database uses a well-researched data model that provides flexibility and scalability on the technical side and adaptability to low-resource settings based on the main features of healthcare systems in such settings. The underlying data model is illustrated in Figure 8.3.

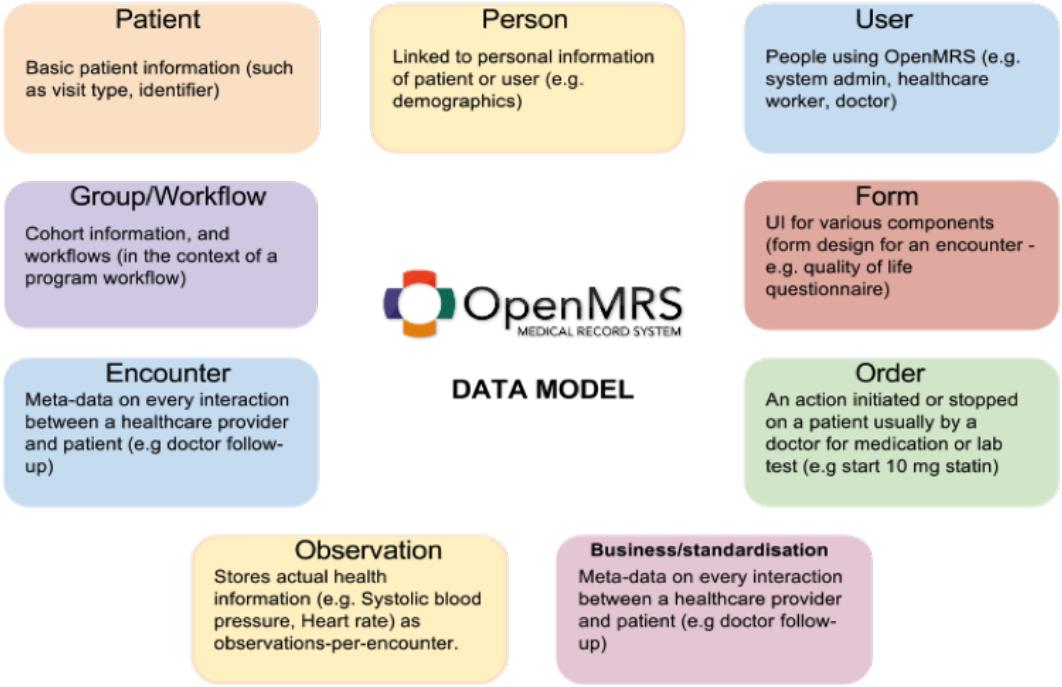


Figure 8.3 Building blocks of the OpenMRS data model in the context of CVD risk management.

Service layer The service layer comprises a comprehensive Application Programming Interface (API) that is a wrapper to the data model and which thereby allows interaction through Java objects. The Spring framework is used in the service layer; for instance, abstract-oriented programming for authentication and logging, and dependency-injection for providing dependencies across components.

User Interface layer The User Interface (UI) layer uses the Spring MVC and is used to provide the Model-View-Controller design pattern, Direct Web Remoting (DWR) for

easy AJAX functionality (where Java objects and methods can call JavaScript objects and methods and vice versa), JSP and JavaScript.

OpenMRS emphasises the standardisation of terminology, whereby coded terms are preferred to free text. This is enabled by the use of a centralised concept dictionary that stores all the clinical terms used in the system (such as those for a particular examination, report, or drug list).

OpenMRS modules A key feature of OpenMRS is modularity, so that developers can add modules to perform new functions (for example, sending a text message from the server or running a scheduler to perform routine tasks), based on interactions with all three layers of the data model. The only core module currently used is the Logic module, which is a service that provides 'logic rules' to encapsulate business rules (e.g. the definition of a 'paediatric' patient). Packages such as Groovy and AngularJS are also used in third-party modules, which are stored in a large module repository.

8.3.2 System architecture and key improvements for SMARThealth RCT

The augmented system architecture for the RCT intervention is described in Figure 8.4. An overview of the sequence of steps is described below. Key improvements are discussed for every step.

1. **Household screening by ASHAs** - Equipped with a Samsung Galaxy 3 tablet with the *SMARThealth-intervention* application, ASHAs are provided with a unique login code when they are registered with the SMARThealth programme. Upon the ASHA entering this code, the tablet's *SMARThealth-intervention* application loads those participants

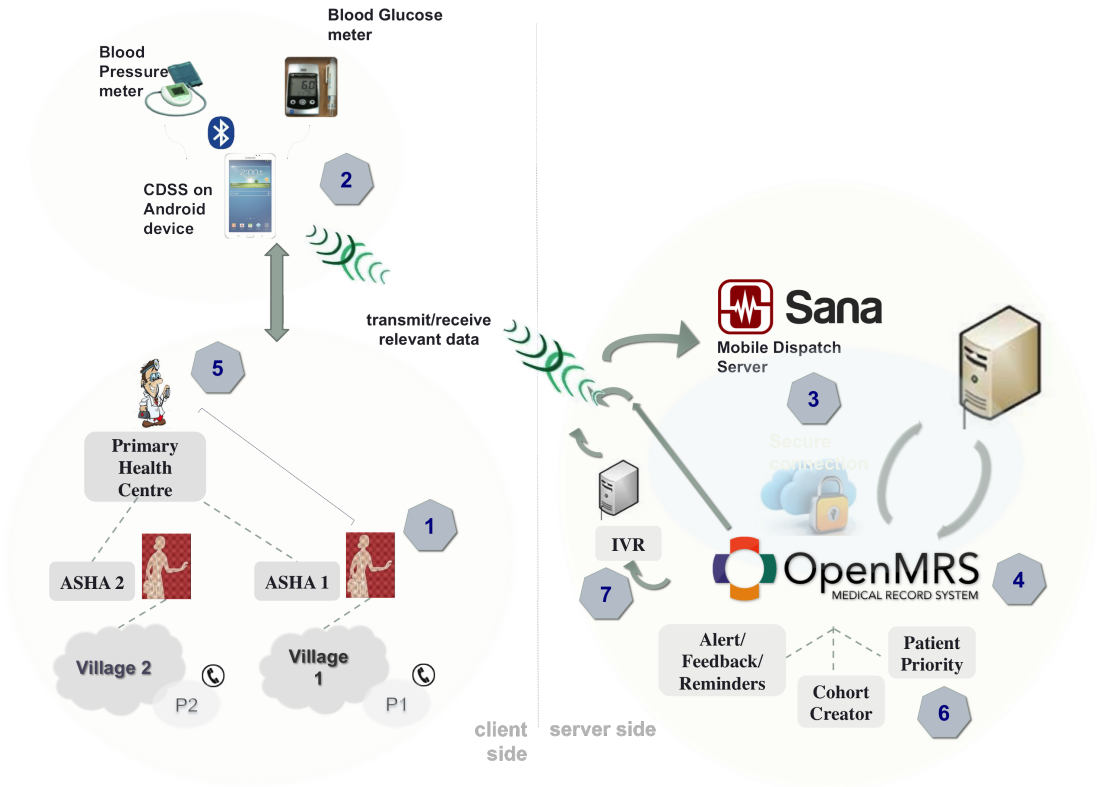


Figure 8.4 mHealth architecture for SMARThealth RCT. The numbers 1 to 7 correspond to the following tasks: 1) Household screening by ASHAs; 2) Decision support offered to ASHAs; 3) Data transmission and storage on the server-side; 4) Server-side data management; 5) Data retrieval by a physician at the PHC; 6) Further server-side processing; and 7) Feedback through an Interactive Voice Response system

residing in the ASHA's locality into her tablet. Hence, ASHAs can perform door-to-door risk assessment covering all 62,254 participants registered during the baseline study. Also, the ASHAs are restricted to screening participants who were already registered during the baseline study. Furthermore, the list of high-risk participants (who can be referred to as our 'patients') identified during the baseline study are not provided to the ASHAs in order to mimic the real-world scenario. The data collected was identical to that of the pilot study (as described in Table 2.1 in Chapter 2) and included the participant's demographics, medical history, family history, treatment history, risk factors (blood pressure, blood glucose), and anthropometric measurements.

2. **Decision support offered to ASHAs** - Upon finishing a risk assessment procedure, the CDSS provides the 10-year risk of developing CVD for the participant as well as point-of-care (POC) decision support on referral and CVD management. This implies that those identified at high risk of CVD by the CDSS are referred to their nearest Primary Health Centre (PHC). high-risk participants, also described as patients, are also presented with a referral card that incorporates a barcode, as illustrated in the next section. The barcode is dynamically associated with the patient's identifier by photographing the barcode in the referral card using the tablet's camera, thereby establishing an association between the barcode and the patient's identifier.

   Data from the Samsung tablets is then uploaded securely to the server. A key improvement of POC recommendations is the incorporation of animations to encourage smoking cessation and maintain healthy diet and adequate physical activity. This is designed to aid the ASHAs to explain the CVD risk recommendations to participants and is shown in Figure 8.5.

3. **Server-side data management** - The mechanism for data processing on the server-side is as follows: data sent by the client in the JavaScript Object Notation (JSON) format is received by the Sana Mobile Dispatch Server (MDS). The MDS, running on the Django framework, subsequently forwards data on to the OpenMRS system. The key roles performed by the MDS are to capture meta-data such as time and frequency of data uploaded from different tablets, to act as a back-up for data storage, and to cache data in case the server running OpenMRS is down. Two key components facilitated by the server-side system are:

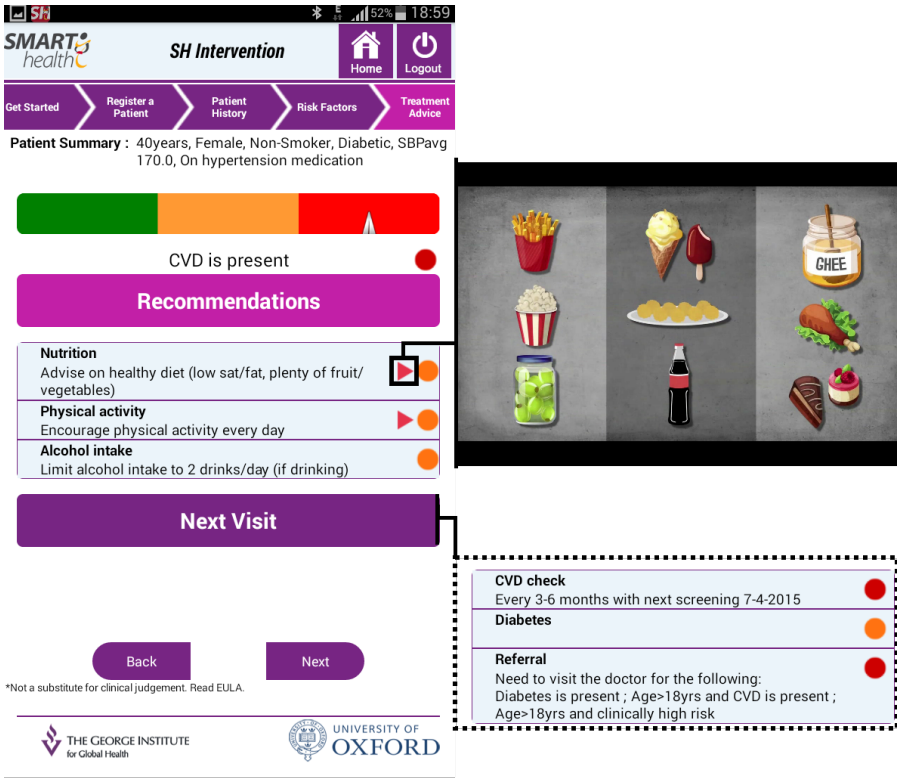


Figure 8.5 Animations implemented in the *SMARThealth-intervention* application for decision support

Enhanced server-side monitoring - An important factor in the success of the sw-RCT is utilisation of the capabilities of local resources and workforce. This is because the effectiveness of the delivered intervention will depend on how well the mHealth system fits into the understanding and the workflow of local health providers. Previous mHealth trials such as the Oxford type-1 diabetes trial have demonstrated the importance of the human element (such as a telehealth nurse) in the intervention loop [213]. In order to build a robust server-side platform that is acceptable and convenient to use for field supervisors, improvements were performed on the server-side web application (previously described in section 4 of in Chapter 3) which assisted visualisation of frequency of uploaded data, and provided a portal for conveniently exporting data from OpenMRS. The following improvements were made: last login times for each ASHA (revealing the frequency of CDSS usage), and a reporting tab providing granular

information on the percentage of high-risk participants (previously recorded at the time of the baseline study) screened in every locality and village.

Security and data storage - Our back-end systems including the OpenMRS and web-application are protected by a self-signed certificate (primarily because no third-party users need to trust the certificate) with a TLS encryption [214]. Additionally, OpenMRS has an internal security model that is based around privileges for particular data types (permission is needed in every respect, be it editing encounters, forms, etc.). OpenMRS also underwent a professional security review (by Aspect Security) in 2008. The *SMARThealth-intervention* application uses tablet-based authentication for data entry and retrieval. Registered tablets can have their data wiped out remotely in the eventuality of theft. The user data that has been synced with the server can be synchronised anytime on any new tablet. De-identified patient data is used for continual synchronisation via proxies such as encounter type, patient identifier, and other encounter-related parameters. Identifiable patient data is used during the time of a first assessment only to register the patient on the server.

4. **Server-side processing** - OpenMRS primarily provides two technical capabilities - (i) the provision to build custom modules to process patient data and automatically schedule continual execution of code and (ii) a RESTful API to facilitate a web service that enables exchange of data for synchronisation between the client-side tablets and server-side medical record system. Three custom modules were developed to process patient information, as described below.

   Cohort creator module - This module forms a cluster of high-risk participants (a 'cohort') handled by each ASHA. It includes only those participants who have been screened by the ASHA until that point in time. Cohorts are periodically

scheduled for update to associate any new high-risk participant data with the respective ASHA. Similarly, clusters of high-risk participants seen by physicians at every PHC are created.

Patient priority module - The ratio of ASHAs to participants in this study is 1:90 in the villages. The design of the sw-RCT means that villages randomised to receive intervention early will lead to an increase in the workload of ASHAs. This is on account of having to manage follow-ups as well as screening new participants. We developed a simple algorithm to prioritise follow-up of patients and help the ASHA manage her schedule. The prioritisation algorithm used a points-based scoring system for high-risk participants who have been screened for CVD risk by the ASHA. The following parameters were used: *whether the participant visited the PHC subsequent to referral by the ASHA*, *whether treatment was prescribed*, *whether self-reported medication adherence was above 75%*, *the change in smoking status*, *whether the BP target (<140/90) was reached*, and *whether the next assessment by an ASHA/physician was imminent in the next 2 weeks*. The scores for all high-risk participants within an ASHA's jurisdiction are sorted in descending order. Subsequently, the top-ranked participants needing attention/follow-up are according to the degree of effectiveness of the ASHA's intervention. For instance, modifiable factors like smoking, adherence to medication (based on participant's self-reporting), and visit to a PHC after referral can be largely controlled, and the ASHA has the potential to influence the participant's behaviour substantially. The priority scores are updated every day through the use of a scheduler in OpenMRS.

Alerts/Reminders - This module processes vital statistics on the use by the ASHA and the physician of the *SMARThealth-intervention*. For instance, parameters such as the last timestamp of an ASHA's login into the tablet, and the number of assessments performed by an ASHA are recorded and stored as an OpenMRS 'attribute'. Attributes are additional information associated with end-users, and can store information that is prone to regular change and does not require historical records. This enables a faster and more convenient retrieval of these parameters through OpenMRS's RESTful API.

5. **Data retrieval by a PHC physician** - high-risk participants, when visiting their PHC physician, are advised to carry their referral cards. This enables the PHC physician to scan the barcode on the referral card using the tablet's camera. This results in automatic display of the participant's previous risk assessment data. In case patients do not carry their referral card, the physician can search for them using their village, gender, and last name. POC decision support on treatment recommendations are offered to the physician. The class of medications suggested (such as five classes of antihypertensives) were proposed by our clinical collaborators in the sw-RCT at the George Institute of Global Health, India. The recommendations follow the National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular diseases & Stroke (NPCDCS) guidelines outlined by the Indian Ministry of Health and Family Welfare [1]. Additionally, these classes of drugs are commonly available in the PHC. The POC decision support for treatment is shown in Figure 8.6. Contraindications on popular class of drugs for BP lowering, lipid lowering, and anti-platelet therapy are also provided. Subsequently, the PHC physician reviews a follow-up care plan and the data is uploaded to the server.

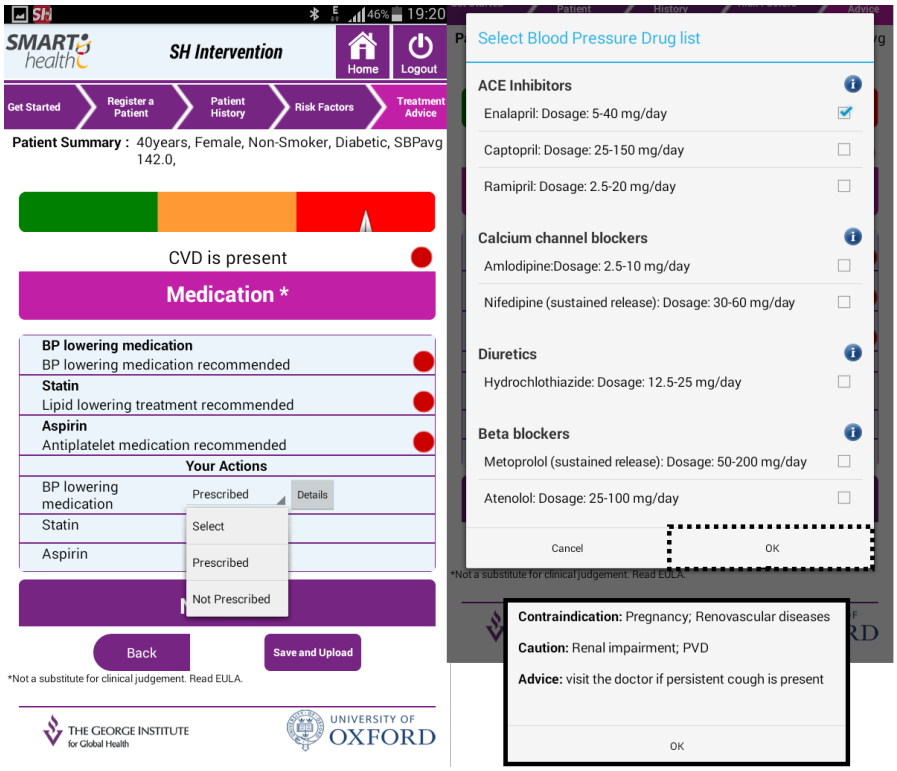


Figure 8.6 POC decision support on treatment offered to a physician. The recommendations are based on the Indian national guidelines (NPCDCS) [1]

6. **Further processing** - Further server side processing is performed as the patient priority module updates the priority scores every day. Two records of the priority score are stored - the previous and current values. The client-side tablet application retrieves a list of patients the ASHA must attend to in the order of decreasing 'priority' score. This guides ASHAs on real-time management of follow-up patients. For effective management of the ASHA's time and workload, as well as to increase frequency of follow-up for each patient including measurement of BP (which is the main outcome of the swc-RCT), a shorter data-collection process for follow-up procedures (also referred to as a 'quick follow-up') has been incorporated into the mobile application. This allows ASHAs to record the following parameters only: BP, medication history, self-reported medication adherence, and visit to a PHC following referral.

7. **IVR feedback** - Rural communities in India receive health information through a multitude of channels that are different from urban areas. In particular, 'word of mouth' is an important channel in rural communities along with health education campaigns, briefings, posters, celebrating awareness weeks and associated workshops, radio, as well as counselling by community health workers. Mobile phones can play a major role in disseminating important health information in rural communities, where access to the internet and media are limited. Following the pilot study described in Chapter 2, we discussed the limitations of using SMS for sending alerts, which include language barriers, spam, low adoption, and delayed service of SMS texts. Therefore, based on metrics derived from the Alerts/Reminders module, an Interactive Voice Response (IVR) system was installed. It primarily automates voice reminders for ASHAs and PHC physicians to follow-up/review high-risk participants. The IVR system was installed by a company contracted by the George Institute of Global Health, India.

8.4 Analysis - snapshot of RCT intervention data

Data from participants who are in the intervention as of 12^{th} September 2015 are chosen for interim analysis here. This encompasses the first 9 months of intervention during which ASHAs performed over 35,195 risk assessment procedures across 12 PHCs and 18 villages. This includes the screening of 27,346 participants, the majority of whom are from the first arm of the stepped wedged cluster. The first phase of the RCT was administered from December 2014 till May 2015 and included 18 villages and 6 PHCs (see Figure 8.1). The second phase commenced in June 2015. Analysis and reporting of trial results were restricted because of concerns that interim analysis may bias the final outcome of the trial. The analysis presented here therefore focuses only the change in SBP for participants who received the

RCT intervention.

At the end of the first phase of the intervention, the 27,346 participants screened covered 65% of the target population. This estimate was based on comparisons with the baseline data described in Chapter 3. high-risk participants screened by the ASHAs included 64% of those identified to be at high risk during baseline screening.

Out of 27,346 participants screened, 6529 were followed up. Four participants for whom follow-up timestamps were not recorded were excluded. The number of participants who have a high risk of CVD is 4650, out of whom 3085 were followed up by a health provider at least once. Therefore a total of 7735 risk assessments were conducted for high-risk participants. The frequency of follow-up assessments for high-risk participants is described in Figure 8.7.

A total of 131 participants were assessed and followed-up five or more times by an ASHA or PHC physician. We shall consider this group of 131 participants for further analysis as they had at least 5 risk assessments performed after receiving the intervention. We shall refer to these as the *subgroup* of participants chosen for analysis. The mean SBP levels (computed as the average of the last two of the three BP measurements acquired from each participant) recorded during each assessment for the subgroup are shown in Figure 8.8. For clarity, these mean SBP levels for each participant will be referred to as simply the SBP levels of the participants. The mean difference between the time duration of first assessment (or screening) and fifth assessment (or follow-up) was 158.28 ± 47.82 days (or over 5 ± 1.5 months).

During the baseline study (that was described in Chapter 3), the mean values of the SBP levels for the subgroup was computed to be 143.01 ± 25.07 mmHg and the mean values of the

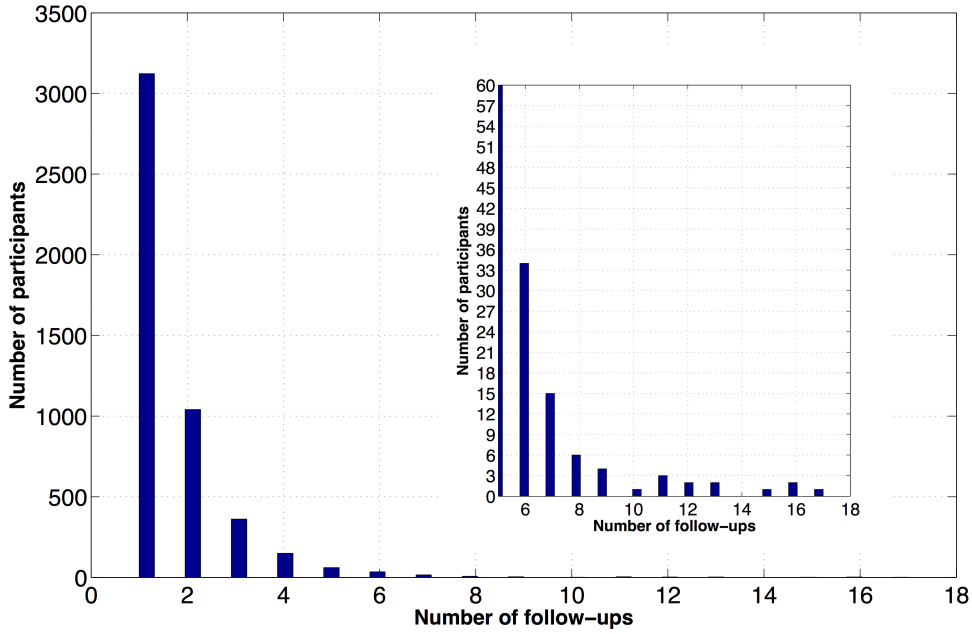


Figure 8.7 The number of risk assessments performed per participant during the intervention. A majority of the participants were assessed once by an ASHA. The number of follow-up procedures greater than 10 was low during the first phase of intervention. However, 131 participants were assessed for CVD risk at least five or more times (shown in the zoomed-in plot), including follow-up procedures.

DBP levels for the subgroup was calculated to be 84±12.34 mmHg. After the baseline study, participants were under 6 months of 'control' or usual care. Subsequently, the participants in our subgroup shifted to receiving the intervention. At the time of the first risk assessment procedure or the first screening by an ASHA during the intervention, the subgroup had largely unchanged mean SBP levels of 143.73 ± 26.85 mmHg while the mean DBP levels were 82±13.39 mmHg. During the fifth assessment, they recorded mean SBP levels of 138.41±23.56 mmHg and mean DBP levels of 80±37.80 mmHg. A Wilcoxon signed rank test, which is a non-parametric hypothesis test for paired data samples was performed. The difference in median SBP levels between the first assessment and fifth assessment was statistically significant (p=0.0097). However, the difference in median DBP levels between

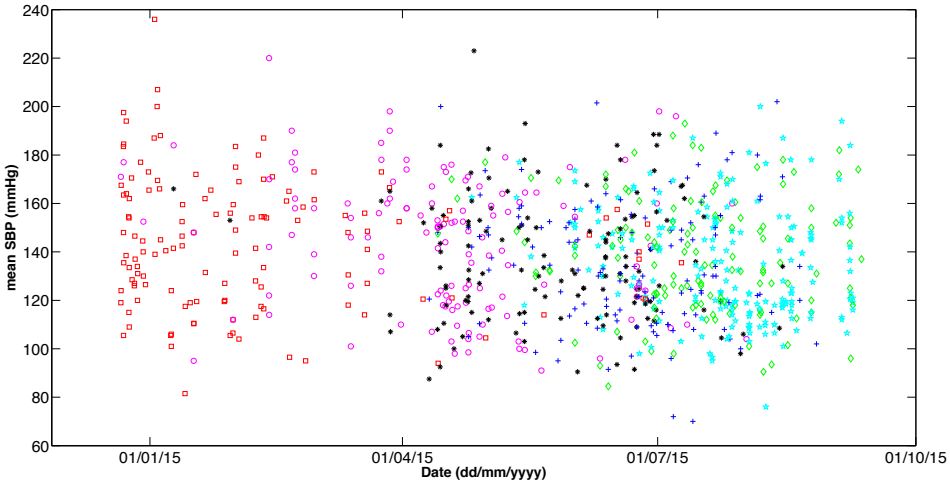
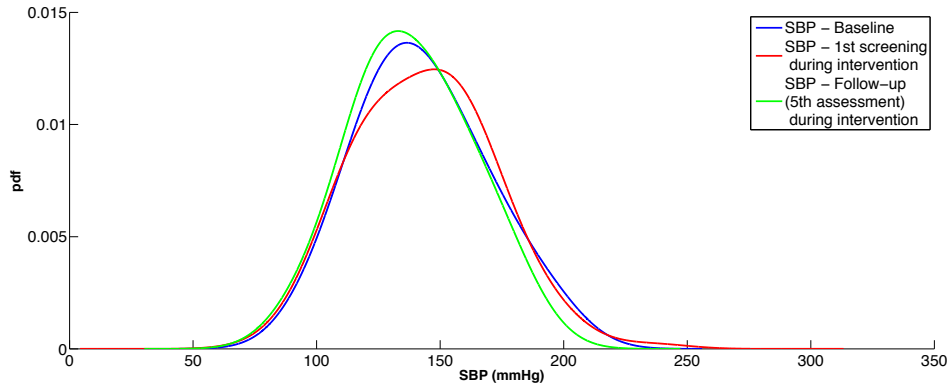


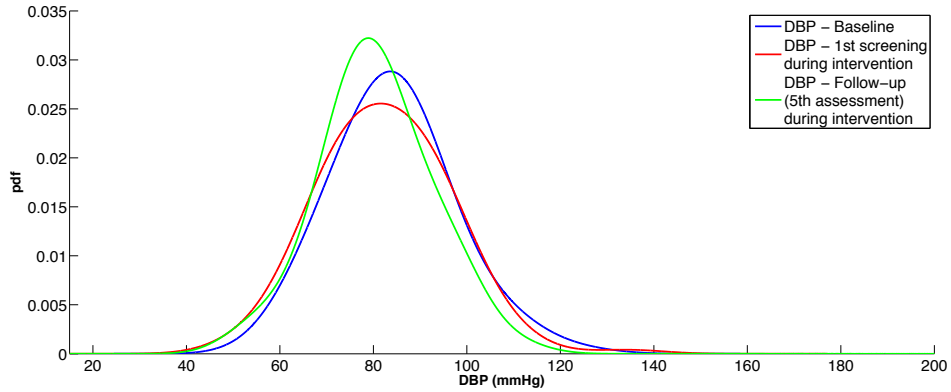
Figure 8.8 Mean SBP levels of participants in the intervention group, who have had at least five risk assessments ($N_{f5} = 131$). The first risk assessment or screening by an ASHA is plotted in red; the second assessment by a PHC physician is plotted in maroon; the third, fourth, and fifth assessments are illustrated in black, blue, and green respectively. Further follow-up assessments are coloured in cyan.

the first assessment and fifth assessment was not statistically significant (p=0.0510).

The distribution of the mean SBP and DBP levels were estimated at three time points - baseline, first assessment (intervention), and fifth assessment (intervention). This is illustrated through the density plots in Figures 8.9a and 8.9b. Silverman's rule was used to estimate the density [126]. The shift in the tail of the distribution for participants with high SBP or DBP levels during the three time points can be observed in Figure 8.9b.



(a) Estimate of the distribution of mean *systolic blood pressure levels*.



(b) Estimate of the distribution of mean *diastolic blood pressure levels*.

Figure 8.9 Estimate of the distributions of the mean SBP and DBP levels at three points of time: at the time of the baseline study; at the time of the first assessment (start of the intervention); and at the time of the fifth assessment. It is clear that the mean SBP remains the same under usual care (during the 6 months of control) and the shift in density for mean SBP during the fifth assessment is evident. The difference between the median SBP levels of the first and fifth assessments was statistically significant (p=0.0097). However, the difference between the median DBP levels of the first and fifth assessments was not statistically significant (p=0.0510).

The difference in SBP levels between successive assessments and the first assessment is shown in Figure 8.10. The median of the difference in SBP levels between successive assessments continues to decrease throughout the intervention period.

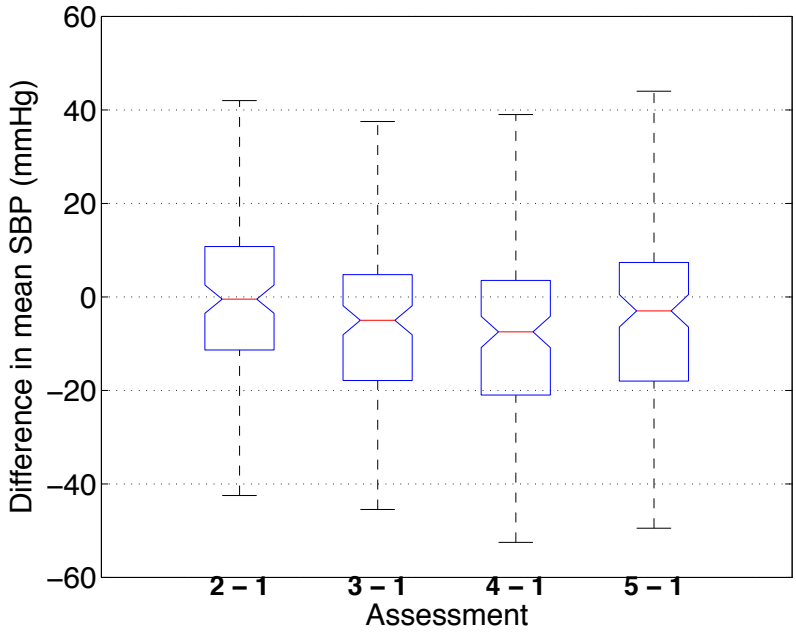


Figure 8.10 Change in SBP level between successive assessments and the first assessment of the intervention. It can be observed that the median difference in SBP levels continues to decrease until the fourth assessment but is marginally higher during the fifth assessment.

Table 8.1 indicates the status for high BP treatment as recorded by the 131 participants who had at least five risk assessments. It is clear that the proportion of participants on high BP treatment continued to rise from the first to fifth assessment.

Table 8.1 Status of treatment for high BP recorded for participants had at least five risk assessments ($N_{f5} = 131$). Participants who maintained the status quo on treatment for high BP either continued to remain on medications or not.

		1^{st} assessment	2^{nd} assessment	3^{rd} assessment	4^{th} assessment	5^{th} assessment
Total	On treatment	74	86	86	88	90
	No treatment	57	44	42	37	34
Change in treatment status (lag)	Newly initiated		13	10	11	13
	Maintained status quo		116	115	114	109
	Discontinued		1	6	5	9
Change in treatment status from first assessment	Newly initiated		13	19	25	26
	Maintained status quo		116	108	104	103
	Discontinued		1	3	1	0

8.5 Conclusion

We have discussed the final improvements performed on the mHealth CDSS to make it production-ready for a RCT in rural India. Data was acquired after 1 year of the RCT with 6 months of intervention and preliminary analysis revealed a statistically significant decrease in median SBP for participants in the intervention group who had at least 5 risk assessments. Upon completion of the RCT, a larger dataset will be available and the findings hold potential to generate evidence for how best to manage CVD risk in resource-poor settings using an mHealth CDSS.

Chapter 9

Summary and future work

9.1 Summary

It is well established that the leading global cause of mortality and morbidity, cardiovascular disease, is more severe in resource-constrained environments such as rural India as introduced in **Chapter 1**. The challenges for health systems seeking to manage CVD in these regions are enormous, and include inadequate services to detect and manage CVD, the epidemiological shift from communicable to chronic disease, changing lifestyles, and the lack of trained healthcare professionals.

The development of a mobile-based, point-of-care clinical decision support system, *SMARThealth* was described in **Chapter 2**. The system could calculate the 10-year risk of CVD in the rural Indian population and help manage high-risk people. An important aspect during the design and development of the system was the use of agile development and user-centred design, which cohesively involved all stakeholders (including the ASHAs, PHC doctors, and community participants), and made the tool suitable for use by ASHAs. This differentiates the *SMARThealth* system from conventional global health projects which 'push' solutions onto the community. During pilot testing with 292 participants, in over

72% of screening procedures performed, the mobile application was found easy to use for that procedure. Users concurred similarly on the usefulness of a graphic risk projection meter, which visualised risk scores, in communicating the meaning of absolute CVD risk reduction to the community participants. In less than 2% of the procedures performed, the ASHAs recorded difficulties with collection of risk factors (such as blood pressure, height and weight, blood glucose). This is vital since ASHAs, who from interface between the public health system and the doors of Indian villages, are provided with an opportunity for capacity enhancement thereby enabling the utilisation of a minimally trained healthcare workforce for managing chronic disease.

Another important aspect from the *SMARThealth* pilot test was to also understand if the system of delivering CVD risk assessment and management could be scaled-up. Two subfactors critical to this are: workflow integration such that frequent use may be possible, as well as to understand user behaviour and uptake. To assess these aspects, a mobile analytics framework enabled quantification of parameters including *system efficiency* (observed through a decrease in median procedure time as more CVD procedures were performed), assessment of manual data entry errors, usefulness of point-of-care management recommendations to ASHAs, as well as an evaluation of how different ASHAs performed. The sub-group analysis sampled ASHAs in three categories: the best performing ASHA, the ASHA with most consistency in performing risk assessments, as well as the ASHA requiring additional attention and/or frequent monitoring. Qualitative examples from the interviews with ASHAs (Appendix C) enabled understanding of barriers to adoption including aspects related to patient's understanding of the management of CVD risk based on the CDSS outputs.

The scale-up of *SMARThealth* is presented **Chapter 3**, where large scale data collection across 54 villages and 18 PHCs was performed as part of a baseline analysis from our target

population. Data was collected from over 62,194 participants, out of whom 9864 or 15.8% of the participants were at high risk of CVD.

This chapter describes in detail how local infrastructure and resources could be utilised for large-scale mHealth data collection in a resource-constrained setting. Important issues relating to handling regions with weak network signal and increasing coverage of data collection through mapping households and respondents were addressed. Minute process improvements helped minimise data loss from failure to assess an eligible participant. Examples include: the use of a web-application which not only connected to an online medical record but also constituted an improved interface (similar to Microsoft Excel) for empowering local field-workers (who were familiar only with the Excel-styled interface) to monitor the data collection process remotely; and the two phase process that took advantage of census data to identify every household in the region and verified the eligibility criteria of all participants from 54 villages.

Two key aspects observed from the baseline data with respect to CVD risk factor prevalence were: the disparity between the number of participants who were hypertensive (40.5%) and those on treatment (19.6%); and that the prevalence of smoking in males was eight times greater than in females. The acquired data was visualised with two techniques: the more recent t-Stochastic Neighbour Embedding technique and a standard non-linear technique using neural networks called Neuroscale. The data points in the low-dimensional space was coloured with CVD risk ranges based on the WHO/ISH low information charts, and the groups formed suggested an unsupervised approach can be used to identify high CVD risk participants.

The major component of the *SMARThealth* system is the CVD risk prediction algorithm. Three widely used risk prediction algorithms, namely the FRS, WHO/ISH charts, and the QRISK2 risk score were compared in **Chapter 4** to investigate how well it performs on data from rural India. The FRS main version under-predicted risk in a rural Indian population as compared to the FRS simplified version with over 20% of the dataset (222 patients) exclusively classified as low risk (0-<10%) by the former. Case studies were presented to show the influence of gender and total cholesterol in CVD risk prediction scores calculated using the WHO/ISH risk prediction charts. In particular, a non-linear interaction between gender and diabetes was observed as exemplified by the fact that females were at higher risk of CVD than males in all cases, except for those patients who have diabetes and who obtained their risk score by specifically using the WHO/ISH high information (HI) chart. The effect of the deprivation score on the QRISK2 algorithm was evaluated at three levels: maximum, minimum, and median deprivation.

In **Chapter 5**, it was found that the choice of LI or HI WHO/ISH risk prediction charts was statistically significant for CVD risk prediction in rural Indian residents (p=0.008;χ^2=7.03) with 155 subjects (or 14.5% of 1066 patients) being given different CVD risk scores by the LI and HI WHO/ISH risk models. Furthermore, the LI and HI WHO/ISH charts disagreed on the clinically relevant classification of 31% of subjects in a sample of 155 high-risk individuals who needed treatment (T_{HR}). Assuming that the HI WHO/ISH risk prediction charts were more accurate than the LI charts (on account of using more information), the LI chart was observed to overpredict risk for T_{HR} patients. Therefore, the inclusion of total cholesterol, which was the only predictor that was excluded in the LI WHO/ISH chart as compared to the HI WHO/ISH chart, was important in risk prediction.

To identify patients for whom the assessment of TC would be beneficial prior to the computation of CVD risk, a parsimonious POC test was developed using established machine learning techniques namely Random Forest, Regularised Logistic Regression, and Support Vector Machines. After feature selection, the POC test developed identified the two main risk factors as being age and SBP, which are routinely collected during risk assessment with the WHO/ISH charts. The analysis performed found good discrimination of the POC test with out-of-sample AUCs of 0.85 (RF), 0.84 (SVM), and 0.82 (RLR). The performance of RF on test data shows that at the maximum F_1 score, a sensitivity of 87% and specificity of 71% were achieved while at the maximum F_3 score, we obtained 91% sensitivity and 62% specificity. An understanding of the differences in risk prediction between the LI and HI models, and adoption of a pre-screening POC test to assess the benefit of a TC test, can aid planning for resource-allocation and estimating the potential saving in costs for large-scale screening programmes.

There is a paucity of data with recorded 10-year outcomes on cardiovascular disease in many lower and middle income countries including India. To compare CVD risk prediction models with known outcomes and to identify highly predictive features for CVD risk that were similar to the *type of inputs* required by the mobile-based CDSS, *SMARThealth*, data from the Framingham study, which has 10-year CVD events recorded, was utilised as described in **Chapter 6**. Eight highly predictive features were identified, and the resulting prediction model that was trained on the Framingham data (referred to as *Model 1*) performed comparably to the two versions of FRS. This was exemplified by higher or equal AUC (0.748±0.03 for *Model 1*; 0.720±0.01 for FRS-3 simple; 0.729±0.01 for FRS-3 main), higher log-likelihood score (-239.99±9.59 for *Model 1*; -251.70±4.23 for FRS-3 simple; -247.04±4.00 for FRS-3 main), and lower Brier score (0.093±0.004 for *Model 1*; 0.097±0.001 for FRS-3 simple; 0.095±0.001 for FRS-3 main) on out-of-sample data from Exam 6 of the

Framingham Offspring study.

The contributions of age and gender on their own could lead to AUCs close to 0.70, as was demonstrated by *Model 1*. However, the mere use of age, or age and gender for CVD risk screening is a complex question with a latency between evidence and action [182]. On application of *Model 1* and the two versions of the FRS to the rural Indian population, lower error rates (MAE: 11.79 and RMSE: 13,59) and better calibration were observed for *Model 1* in comparison to the two versions of FRS, assuming that the HI WHO/ISH risk model can be considered as the benchmark. The possibility of recalibrating of *Model 1* for rural India was introduced.

The lack of recorded end outcomes in rural India prompted the use of an unsupervised approach to identify high-risk patients as described in **Chapter 7**. Hierarchical clustering was applied to the rural Indian dataset, and to data from the Framingham study. Low CVD risk and high CVD risk clusters were found when $\hat{K}$=2, but also clusters with intermediate risk when $\hat{K}$=4. It was found that the cluster compositions and $\hat{K}$ were consistent across both datasets (rural India and Framingham). Thus, unsupervised clustering offers an alternative approach to identifying groups of high-risk patients in rural India.

Previous research has shown the lack of scaling-up in mHealth studies, with only 9 mHealth randomised controlled trials reported in LMICs. The improvements needed to build a production-ready *SMARThealth* CDSS for evaluation in a stepped wedged RCT were discussed in **Chapter 8**. Notable contributions to the CDSS included the design of a mechanism for bi-directional synchronisation with OpenMRS for efficient uploading and retrieval of data in a POC setting. By efficient, we mean minimising the load transactions during every step of the risk assessment procedure and faster processing (e.g. facilitating key

variables such as patient ID, priority score to be defined as 'attributes' on OpenMRS) .

No data is collected regarding event rates from the control groups in the RCT. However, it has to be kept in mind that in a stepped wedged cluster RCT, every control group eventually moves to intervention. In such a case, event rates are captured by the questions on medical history of the patient that is collected upon every ASHA follow-up. If the patient has visited either private or public clinic, this is recorded. This is a superficial way of capturing event rates however, given the focus of the RCT, this may be the best that can be achieved. The earlier APHRI study was specifically run to collect data from morality registers on CHD. Because of the diversity in private practice, and a common practice of visiting urban hospitals if a villager deemed his condition to be important, it is difficult to perform data collection on events. This is why no major long-term study has been in place yet. Provision of information technology-driven infrastructure, such as the verbal autopsy developed at UCL may help in easing the data collection process [215]. However, specific training needs to be given and the efficacy and efficiency of such a tool is difficult to be determined without proper investigation in the local context.

The thesis concludes with a presentation of the preliminary data collected during the first year of the RCT, with high-risk patients having at least five follow-ups selected for analysis. A statistically significant reduction in median blood pressure was observed for this patient group (p=0.0097). It was also clear that the proportion of patients on treatment for high BP continued to rise from the first assessment to fifth follow-up. On completion of the RCT, a larger dataset will be available to validate the effectiveness of *SMARThealth* to guide CVD management. The CDSS has the potential to generate evidence on how best to manage CVD risk in rural India.

One of the other key contributions in this thesis is the open source software for mHealth research which was described in **Chapters 2**, **3**, and **8**. To enable interoperability amongst mHealth systems and avoid duplication of effort (as well as the risk of fragmentation), a standard electronic health record system, OpenMRS (version 1.9) [46], and a standard mHealth platform, namely the Sanamobile system (version 2.0) [48][49] were integrated into the CDSS. Three notable contributions to these open-source platforms were:

Addition of a Sanamobile Client Library Sanamobile is an open source telemedicine platform [48]. The work described in this thesis makes two novel contributions to the development of this open-source platform: (a) the development of a novel client side Java library; and (b) the addition of new functionality. Specifically, the first contribution was the development of a client-side library called *sanaClientLib*, that provided an application programming interface for allowing any third party mobile application to connect to a server-side electronic health record through the Sana Mobile Dispatch Server. The library provided all existing features of the Sanamobile client application and allowed the user the choice of any Graphical User Interface. More specifically, it offered flexibility and customisation, such as usage of any database schema, thereby encouraging modularity. The second contribution was the improvement of the Sana Mobile Dispatch Server, whereby support for OpenMRS 'patient attributes' and encounters were added. This allowed faster access and retrieval of patient data at point-of-care through storing frequently used variables as attributes rather the existing method that is encounter-based.

OpenMRS SMARThealth module Another addition to the sever-side OpenMRS [46] system was a scheduled patient management routine for healthcare workers. An algorithm (see Section 8.3.2) was created that utilised a points-based scoring system in order to rank and prioritise all patients in the medical record system for follow-up care. The score was based on their medication adherence, modifiable risk factors (such as

smoking status and blood pressure levels), and compliance to routine CVD screening and referrals. The ASHA then advises either referral/re-visit to a physician, and/or guidance (in the form of animations) for smoking cessation or importance of medication adherence. This algorithm is distributed to run on individual cohorts, then are the subset of patients that belong to one ASHA and fall within the jurisdiction of one primary care centre. This helps the ASHA to allocate priority and focus on patients who need attention, thereby increasing the efficacy of the intervention.

Web-application compatible with OpenMRS A key improvement was the development of a web-application interoperable with the OpenMRS database schema. It provided four advantages. First, it offered a convenient graphical user interface that was similar to Microsoft Excel, software with which the local workforce was familiar. Second, the web-application enabled regular monitoring of assessments performed by each ASHA through simple metrics such as frequency of assessments within a time period. Third, it provided a way for convenient filtering and exporting of data without affecting the OpenMRS server. Finally, visualisation of frequency of data collected from each village was possible.

The improvements made to the existing Sanamobile and OpenMRS platforms were both generic as well as specific to CVD risk assessment and management. Though interest in mHealth has been increasing [216], the lack of coordinated and tailored delivery within an organized framework persists as a major obstacle to progress. *Ad hoc*, isolated implementations may also result in insufficient considerations for the needs of the population. This is highlighted by mHealth systems which, when indiscriminately applied to weak healthcare infrastructures, are likely to end up with a transitory or insignificant impact [21]. Our incremental approach to building mHealth systems using standardised frameworks may therefore inform better strategies for scaling up and effectiveness, as well as reduce the risk of fragmentation of purpose and redundancy of effort.

9.2 Future work

The wealth of data currently being acquired as part of the large-scale RCT in Andhra Pradesh provides interesting directions for future work.

Evaluation of RCT outcomes The primary outcome from the 2-year RCT is to determine whether high-risk individuals can achieve adequate control of their blood pressure when subjected to the intervention used in the trial.

As detailed in chapter 5, BP levels in Andhra Pradesh are largely identical to those reported in urban India [149][150] for similar age groups, which suggests that the Andhra Pradesh region is at an advanced stage of transition. Longitudinal data from those above 40 years of age but who are not at high CVD risk will also be available at the end of the RCT. This is important as it facilitates the creation of prediction models for those likely to develop hypertension within a given time period in that region. Although we may be able to predict only short-term risk of hypertension (less than 2-years), the outcome will be crucial as no known risk models are available in rural India as yet.

In the literature, models for the prediction of hypertension have been developed from the Framingham heart study [217]. The Framingham hypertension model uses a parametric accelerated failure time model (described in chapter 4), the premise of which states that all patients have the same shape of survival curve but that some move faster or slower as per their co-variates. The covariates used in the Framingham hypertension model, namely age, gender, SBP, DBP, BMI, family history of hypertension, and smoking status have been collected in the RCT in Andhra Pradesh as well. The FRS hypertension model was developed on white individuals who did not have diabetes nor high blood pressure at baseline. Since the RCT in Andhra Pradesh will record the incidence of hypertension, the FRS hypertension

model could be recalibrated, similar to the approach of recalibrating the Framingham CVD risk model to rural India as detailed in Chapter 6.

Information on the number of participants who were smoking at baseline and through the duration of the RCT is known. Valuable data on the number of participants who quit smoking and those who resumed after quitting once could therefore be obtained. This ties in closely with the animations embedded in the SMARThealth RCT application (as described in Chapter 8) because whenever an ASHA follows up with a high risk patient, she is trained to impact the modifiable factors of CVD, mainly smoking status and diet. Hence the secondary trial outcomes include mean change in CVD risk factors such as smoking status, as well as diabetes, quality of life, number of CVD events, and process outcomes (such as referrals to a physician).

Formulating deprivation scores from the target population In Chapter 4, a comparison of three major CVD risk algorithms on data from rural India was performed. The QRISK2 algorithm was tested with three levels of deprivation score. The original formulation for the deprivation index was based on four variables obtained from census data specific to the UK: unemployment (lack of material resources and insecurity), overcrowding (material living conditions), lack of owner occupied accommodation (a proxy indicator of wealth) and lack of car ownership (a proxy indicator of income). The West Godhavari community in rural Andhra Pradesh, which is our target population, is a relatively well-off community in comparison to southern rural India (who are in turn more well-off than counterparts in northern India). People who are employed in businesses (aqua-culture) and most zamindars (or landlords who also own vast amounts of farmland) have considerable wealth whilst those involved in labour-intensive jobs (those who do the work in the field) are actually impoverished. A multitude of data will be collected in the RCT, such as quality of life scores, location, occupation, education levels, along with demographic information including

household size, and household income. These variables from the RCT will also include longitudinal information, using which interesting insights on deprivation may be gained. Formulation of a specific deprivation index may then be used for recalibrating the QRISK2 to the Indian population, perhaps through the approach outlined in Chapter 6.

Extending *SMARThealth* to other resource-constrained settings One of the key features in the design of the production-ready SMARThealth system is the system architecture. Not only is it adapted to a resource-constrained setting (e.g. a robust mechanism for bi-directional data synchronisation and fail-proof uploading), but it also encompasses elements of modularity and interoperability. Key constituent elements of the client-side Android application are shown in Figure 9.1.

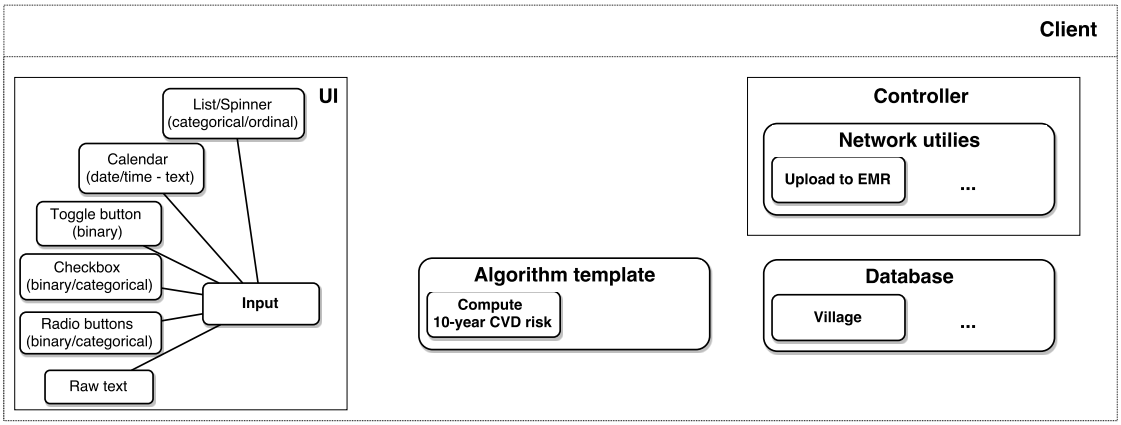


Figure 9.1 Key constituent elements of the client-side Android application. The 'plug-and-play' architecture implies that any desired input or algorithm can be added or removed with minimal modification.

Provision for collecting different input types (e.g. list, multiple choice radio buttons, raw text) and subsequent data pre-processing (e.g checking range of input entry) are enabled. Only the input questions therefore require modification. Once data has been entered, there is a uniform pipeline to process it through to the 'algorithm template' of the SMARThealth system. Any function can be applied to the input data, which can be added in the form of a

module to the algorithm template.

This 'plug-and-play' mechanism is useful for extension of the system to other areas of chronic disease. For instance, the SMARThealth system has recently been used to identify mental disorders in rural Andhra Pradesh. To achieve this, modifications may include selection of the desired questions with an appropriate input type (such as radio buttons) and an algorithm template 'module'. Since the user interface has been customised to resource-constrained settings, it is thought that extension of the SMARThealth system to similar settings could also be achieved.

Non-parametric clustering Non-parametric clustering techniques (such as the Chinese restaurant process) offer a method for finding the optimal number of clusters, with the theoretical possibility of adding an infinite amount of clusters. In addition robustly handling different data types (categorical, numerical, ordinal, and string) is an evolving subject of research and could be investigated in greater detail.

The work of Professor De Leeuw at UCLA in the area of multivariate analysis is particularly relevant. Notably, the R package 'aspect' comprises various functions for optimal scaling [218]. Given a set of questions with responses that are categorical (e.g. say with 3 inputs, namely level of education, marital status, occupation), each category has a resultant transformation which is obtained by maximisation of a particular 'aspect' of the correlation matrix (such as the sum of the first p eigenvalues or the sum of correlations to the power m). Upon transformation, the categorical variables can be conveniently used for analysis with techniques such as PCA.

Building CVD risk prediction models that adapt to new data This thesis outlined CVD risk prediction models based on survival analysis techniques, and demonstrated a machine learning approach to prediction using cross-sectional data in Chapter 6. Although the latter incorporates techniques that capture non-linear interactions, longitudinal information and the subsequent handling of censored variables, which are possible with the former, were not considered. Despite the wide applicability of the Cox regression model for risk prediction, time-varying information of risk factors from patients cannot be used while testing. Future work will include a Bayesian approach to modelling risk, whereby the risk of test patients can be updated as new data is received.

Bibliography

[1] Director General of Health Services, Ministry of Health and Family Welfare, Government of India. Indian national programme for prevention and control of diabetes, cardiovascular diseases, and stroke. 2009 (last accessed: 03-12-14). Available at http://www.searo.who.int/india/topics/cardiovascular_diseases/NCD_Resources_CVD_RISK_MANAGEMENT_BOOKLET.pdf?ua=1.

[2] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, R. Minhas, A. Sheikh, and P. Brindle. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, 336:1475–82, 2008. Available at http://www.ncbi.nlm.nih.gov/pubmed/18573856.

[3] Clara Chow, Magnolia Cardona, P. Krishnam Raju, Srinivas Iyengar, Akamshetty Sukumar, Ravi Raju, Sam Colman, P. Madhav, Rama Raju, K. Srinath Reddy, David Celermajer, and Bruce Neal. Cardiovascular disease and risk factors among 345 adults in rural India- the Andhra Pradesh Rural Health Initiative. *Int J Cardiol*, 116(2):180 – 185, 2007. Available at http://www.sciencedirect.com/science/article/pii/S0167527306004438.

[4] World Health Organisation. Cardiovascular diseases (CVDs) Fact Sheet. 2011 (last accessed: 21-10-15). Available at http://www.who.int/mediacentre/factsheets/fs317/en/index.html.

[5] Thomas A Gaziano. Reducing the growing burden of cardiovascular disease in the developing world. *Health affairs*, 26(1):13–24, 2007.

[6] Gerard La Forgia and Somil Nagpal. *Government-sponsored health insurance in India: Are you covered?* World Bank Publications, 2012.

[7] Kakoli Roy and David Hill Howard. Equity in out-of-pocket payments for hospital care: Evidence from india. *Health Policy*, 80(2):297 – 307, 2007. Available at http://www.sciencedirect.com/science/article/pii/S0168851006000728.

[8] The Economist Intelligence Unit. Industry report; healthcare: India. July 2014 (last accessed: 11-10-15).

[9] J.M. Swaminathan. Indian economic superpower: fiction or future? *World Scientific Publishing Company*, 2009.

[10] Insurance Regulatory and Development Authority (IRDA). Health insurance for everyone. XIII(1), 2015.

[11] Vikram Patel, Somnath Chatterji, Dan Chisholm, Shah Ebrahim, Gururaj Gopalakrishna, Colin Mathers, Viswanathan Mohan, Dorairaj Prabhakaran, Ravilla D Ravindran, and K Srinath Reddy. Chronic diseases and injuries in india. *The Lancet*, 377(9763):413–428, 2011.

[12] Khurana S Kumar H for the Indian Council of Medical Research Shah B KN, Menon GR. Assessment of burden of non-communicable diseases. 2010 (last accessed: 29-07-11). Available at http://www.whoindia.org/EN/Section20/Section306_1025.htm.

[13] Ramadoss A Reddy KS SB, Varghese C. Responding to the threat of chronic diseases in india. *Lancet*, 366:1744–49, 2005.

[14] PricewaterhouseCoopers. Healthcare in India: Emerging market report. Technical report, PricewaterhouseCoopers, 2007.

[15] Venkatesh S Satpathy SK. Human resources for health in india's national rural health mission: Dimension and challenges. *Regional Health Forum*, 10:29–37, 2006.

[16] World Health Organization. WHO India - Origin and Evolution of Primary Care in India. (last accessed: March 9th, 2012). Available at http://whoindia.org/LinkFiles/Health_Systems_Development_Primary_Health_Care_Origin_and_Evolution_.pdf.

[17] Dileep Mavalankar and Kranti Suresh Vora. The changing role of auxiliary nurse midwife (ANM) in India: implications for maternal and child health (MCH). *Indian Institute of Management, Ahmedabad*, 2008. Available at http://openlibrary.org/books/OL18591322M/The_changing_role_of_auxiliary_nurse_midwife_(ANM)_in_India.

[18] Robert Istepanian, Swamy Laxminarayan, and Constantinos S Pattichis. M-health: Emerging mobile health systems. *M-Health: Emerging Mobile Health Systems, Edited by R. Istepanian, S. Laxminarayan, and CS Pattichis. Berlin: Springer, 2006.*, 1, 2006.

[19] International Telecommunication Union. ICT Facts and Figures features, (last accessed: 15-04-13). Available at http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.

[20] World Health Organization. mHealth: New horizons for health through mobile technologies. *Global Observatory for eHealth series - Volume 3*, 2011. Available at http://www.who.int/goe/publications/goe_mhealth_web.pdf.

[21] M.Tomlinson, M.J.Rotheram-Borus, L.Swartz, and A.C.Tsai. Scaling Up mhealth: Where is the Evidence? *PLoS Med*, 10, 2013.

[22] PriceWaterhouseCoopers. Report on health infrastructure and services financing in india. 2012. Available at http://www.pwc.in/en_IN/in/assets/pdfs/publications-2012/healthcare_financing_report_print.pdf.

[23] Joaquin Andres Blaya. *Developing, implementing, and evaluating tuberculosis laboratory information systems for resource-poor settings*. PhD thesis, 2008.

[24] A.J. Farmer, O.J. Gibson, C. Dudley, K. Bryden, P.M. Hayton, L. Tarassenko, and A. Neil. A randomized controlled trial of the effect of real-time telemedicine support on glycemic control in young adults with type 1 diabetes (isrctn 46889446). *Diabetes Care*, 28(11):2697–2702, 2005. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-30944453383&partnerID=40&md5=4ec21dad94f815fc97adc3f6a1251c83.

[25] J.Turner, M.Larsen, O.Gibson, H.Simpson, L.Tarassenko, A.Neil, and A.Farmer. A telehealth system to optimise insulin titration in primary care for patients with type 2 diabetes. *Diabetic Medicine*, 26:147, 2009.

[26] J.Turner, M.Larsen, L.Tarassenkoa, A.Neil, and A.Farmer. Implementation of telehealth support for patients with type 2 diabetes using insulin treatment: an exploratory study. *Inform. Prim. Care*, 17:47–53, 2009.

[27] L. Mbuagbaw, L. Thabane, P. Ongolo-Zogo, R.T. Lester, E.J. Mills, M. Smieja, L. Dolovich, and C. Kouanfack. The Cameroon Mobile Phone SMS (CAMPS) Trial: A randomized trial of text messaging versus usual care for ddherence to Antiretroviral therapy. *PLoS ONE*, 7(12), 2012. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-84870832143&partnerID=40&md5=037074c02980aeb22da7354f9109e37d.

[28] S.A. Stoner and C.S. Hendershot. A randomized trial evaluating an mHealth system to monitor and enhance adherence to pharmacotherapy for alcohol use disorders. *Addiction science & clinical practice*, 7(1):9, 2012. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-84873740937&partnerID=40&md5=5959f4cb93507e5184d497e0ee499462.

[29] S. Githinji, S. Kigen, D. Memusi, A. Nyandigisi, A.M. Mbithi, A. Wamari, A.N. Muturi, G. Jagoe, J. Barrington, R.W. Snow, and D. Zurovac. Reducing stock-outs of life saving malaria commodities using mobile phone text-messaging: SMS for Life study in Kenya. *PLoS ONE*, 8(1), 2013. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-84872512578&partnerID=40&md5=a0bbba00ac673c93ab45e0cc39c532eb.

[30] J. Anhoj and C. Moldrup. Feasibility of collecting diary data from asthma patients through mobile phones and SMS (short message service): Response rate analysis and focus group evaluation from a pilot study. *Journal of Medical Internet Research*, 6(4), 2004. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-14044249442&partnerID=40&md5=6c0711ac9ab1f294adff3a59090fa77a.

[31] L.E. Burke, M.A. Styn, S.M. Sereika, M.B. Conroy, L. Ye, K. Glanz, M.A. Sevick, and L.J. Ewing. Using mhealth technology to enhance self-monitoring for weight loss: A randomized trial. *American Journal of Preventive Medicine*, 43(1):20–26, 2012. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-84862638709&partnerID=40&md5=aa3906452c123c64fbbdf2c53aa9ce5c.

[32] R. Maddison, R. Whittaker, R. Stewart, A. Kerr, Y. Jiang, G. Kira, K.H. Carter, and L. Pfaeffli. Heart: Heart exercise and remote technologies: A randomized controlled trial study protocol. *BMC Cardiovascular Disorders*, 11, 2011. Available

at http://www.scopus.com/inward/record.url?eid=2-s2.0-79959282668&partnerID=40&md5=475c547b16d93b722d147d6ea81622de.

[33] D.J. Vidrine, F.E. Fletcher, H.E. Danysh, S. Marani, J.I. Vidrine, S.B. Cantor, and A.V. Prokhorov. A randomized controlled trial to assess the efficacy of an interactive mobile messaging intervention for underserved smokers: Project action. *BMC Public Health*, 12(1), 2012. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-84865304572&partnerID=40&md5=1caba1fbd8174b127336ec4c20d95846.

[34] Karin Källander, James K Tibenderana, Onome J Akpogheneta, Daniel L Strachan, Zelee Hill, Augustinus HA ten Asbroek, Lesong Conteh, Betty R Kirkwood, and Sylvia R Meek. Mobile health (mhealth) approaches and lessons for increased performance and retention of community health workers in low-and middle-income countries: A review. *J Med Internet Res*, 15(1), 2013.

[35] Cheick-Oumar Bagayoko, Marie-Pierre Gagnon, Diakaridia Traoré, Abdrahamane Anne, Abdel K Traoré, and Antoine Geissbuhler. E-health, another mechanism to recruit and retain healthcare professionals in remote areas: lessons learned from equireshus project in mali. *BMC medical informatics and decision making*, 14(1):120, 2014.

[36] David Peiris, Devarsetty Praveen, Claire Johnson, and Kishor Mogulluru. Use of mhealth systems and tools for non-communicable diseases in low-and middle-income countries: a systematic review. *Journal of cardiovascular translational research*, pages 1–15, 2014.

[37] World Health Organization/International Society for Hypertension. Pocket guidelines for Assessment and Management of Cardiovascular Risk with WHO/ISH Cardiovascular Risk Prediction Charts for WHO epidemiological Sub-regions SEAR-D. 2013.

[38] World Health Organization. Prevention of Cardiovascular Disease: Guidelines for Assessment and Management of Cardiovascular Risk. Geneva 2007. Available at http://www.who.int/cardiovascular_diseases/guidelines/Full%20text.pdf.

[39] Ministry of Health and Family Welfare, Government of India. National Programme for Prevention and Control of Diabetes, Cardiovascular Disease, and Stroke:Guidelines on Assessment and Management of Cardiovascular Risk for Medical Officers. 2009 (Last accessed: 01-07-14). Available at http://www.searo.who.int/india/topics/cardiovascular_diseases/NCD_Resources_Training_module_for_NPDCS_for_health_workers.pdf.

[40] Bashar Nuseibeh and Steve Easterbrook. Requirements engineering: A roadmap. In *Proceedings of the Conference on The Future of Software Engineering*, ICSE '00, pages 35–46, New York, NY, USA, 2000. ACM. Available at http://doi.acm.org/10.1145/336512.336523.

[41] Betty H. C. Cheng and Joanne M. Atlee. Research directions in requirements engineering. In *2007 Future of Software Engineering*, FOSE '07, pages 285–303, Washington, DC, USA, 2007. IEEE Computer Society.

[42] Barry W. Boehm. A spiral model of software development and enhancement. *Computer*, 21(5):61–72, 1988.

[43] Lan Cao and Balasubramaniam Ramesh. Agile requirements engineering practices: An empirical study. *Software, IEEE*, 25(1):60–67, 2008.

[44] United nations document on the the registration of births and deaths act, 1969, india. 2014 (last accessed: 01-09-14). Available at unstats.un.org/unsd/dnss/docViewer.aspx?docID=1733.

[45] TH Westhoff, S Schmidt, W Zidek, and M Van der Giet. Validation of the Stabil-O-Graph blood pressure self-measurement device. *J Hum Hypertens*, 22(3):233–235, 2007. Available at http://www.nature.com/jhh/journal/vaop/ncurrent/full/1002287a.html.

[46] Benjamin A Wolfe, Burke W Mamlin, Paul G Biondich, Hamish SF Fraser, Darius Jazayeri, Christian Allen, Justin Miranda, and William M Tierney. The OpenMRS system: collaborating toward an open source EMR for developing countries. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1146. American Medical Informatics Association, 2006.

[47] Gari D Clifford, Joaquin A Blaya, Rachel Hall-Clifford, and Hamish SF Fraser. Medical information systems: A foundation for healthcare technologies in developing countries. *Biomed Eng Online*, 7(1):18, 2008.

[48] SANA mobile - A remote end-to-end medical diagnostics platform. pages –, 2011 (last accessed: 04-08-11). Available at http://www.sanamobile.org.

[49] Leo Anthony Celi, Luis Sarmenta, Jhonathan Rotberg, Alvin Marcelo, and Gari Clifford. Mobile care (moca) for remote diagnosis and screening. *J Health Inform Dev Ctries*, 3(1):17, 2009.

[50] Gerald Albaum. The likert scale revisited. *J Market Res Soc*, 39:331–348, 1997.

[51] A. Raghu, D. Praveen, D. Peiris, L. Tarassenko, and G. Clifford. Lessons from the evaluation of a clinical decision support tool for cardiovascular disease risk management in rural india. *UNESCO Tech4Dev 2014 conference, Lausanne, Switzerland*, 2014.

[52] Alain Abran, Adel Khelifi, Witold Suryn, and Ahmed Seffah. Usability meanings and interpretations in ISO standards. *Software Qual J.*, 11(4):325–338, 2003.

[53] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 345–352, New York, NY, USA, 2000. ACM.

[54] Clive Nancarrow and Ian Brace. Saying the "right thing": Coping with social desirability bias in marketing research. *Bristol Business School Teaching and Research Review*, 3(11), 2000.

[55] Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. *The psychology of survey response*. Cambridge University Press, 2000.

[56] Rohina Joshi, Clara K Chow, P Krishnam Raju, Rama Raju, K Srinath Reddy, Stephen MacMahon, Alan D Lopez, and Bruce Neal. Fatal and nonfatal cardiovascular disease and the use of therapies for secondary prevention in a rural region of india. *Circulation*, 119(14):1950–1955, 2009. Available at http://circ.ahajournals.org/content/119/14/1950.short.

[57] N Gusi, PR Olivares, and R Rajendram. The EQ-5D health-related quality of life questionnaire. pages 87–99, 2010.

[58] The IPAQ Group. Guidelines for data processing and analysis of the international physical activity questionnaire (IPAQ) - short form. 2004.

[59] Brian A Primack. The who-5 wellbeing index performed the best in screening for depression in primary care. *Evidence Based Medicine*, 8(5):155–155, 2003.

[60] Open Medical Record System(OpenMRS) for developing countries, (last accessed: 21-10-15). Available at http://openmrs.org.

[61] Managing person attributes, openmrs wiki. 2015 (last accessed: 30-07-15). Available at https://wiki.openmrs.org/display/docs/Managing+Person+Attribute+Types,2014.

[62] OpenMRS Wiki page. Reporting Module - OpenMRS. (last accessed: 30-07-15). Available at https://wiki.openmrs.org/display/docs/Reporting+Module.

[63] American Diabetes Association et al. Standards of medical care in diabetes—2010. *Diabetes care*, 33(Supplement 1):S11–S61, 2010.

[64] Aram V Chobanian, George L Bakris, Henry R Black, William C Cushman, Lee A Green, Joseph L Izzo, Daniel W Jones, Barry J Materson, Suzanne Oparil, Jackson T Wright, et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension*, 42(6):1206–1252, 2003.

[65] Maria Cristina Ferreira De Oliveira and Haim Levkowitz. From visual data exploration to visual data mining: a survey. *Visualization and Computer Graphics, IEEE Transactions on*, 9(3):378–394, 2003.

[66] Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. A tour through the visualization zoo. *Commun. Acm*, 53(6):59–67, 2010.

[67] Mark Richardson. Principal component analysis. 2009 (last accessed: 21-10-15). Available at http://people.maths.ox.ac.uk/richardsonm/SignalProcPCA.pdf.

[68] Cariad Chester and Holden T Maecker. Algorithmic tools for mining high-dimensional cytometry data. *The Journal of Immunology*, 195(3):773–779, 2015.

[69] Laurens JP van der Maaten, Eric O Postma, and H Jaap van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41):66–71, 2009.

[70] Joseph B Kruskal and Myron Wish. Multidimensional scaling. 11, 1978.

[71] Lawrence K Saul and Sam T Roweis. An introduction to locally linear embedding. Available at https://www.cs.nyu.edu/~roweis/lle/papers/lleintro.pdf(lastaccessed: 21-10-15).

[72] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[73] John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, (5):401–409, 1969.

[74] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[75] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 833–840, 2002.

[76] Laurens Maaten. Learning a parametric embedding by preserving local structure. In *International Conference on Artificial Intelligence and Statistics*, pages 384–391, 2009.

[77] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[78] David Lowe and Michael E Tipping. Neuroscale: novel topographic feature extraction using rbf networks. *Advances in Neural Information Processing Systems*, pages 543–549, 1997.

[79] Small Ministry of Micro and Medium Enterprises. Government of India. Brief industrial profile of west godhavari district, andhra pradesh. (last accesssed: 29-07-15). Available at http://dcmsme.gov.in/dips/amended%20West%20Godavari.pdf.

[80] Jost B Jonas, Vinay Nangia, Marcella Rietschel, Torsten Paul, Prakash Behere, and Songhomitra Panda-Jonas. Prevalence of depression, suicidal ideation, alcohol intake and nicotine consumption in rural central india. the central india eye and medical study. 2014.

[81] Donald M. Lloyd-Jones. Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation*, 121(15):1768–1777, Apr 2010. Available at http://dx.doi.org/10.1161/CIRCULATIONAHA.109.849166.

[82] Evaluation Expert Panel on Detection et al. Executive summary of the third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii). *Jama*, 285(19):2486, 2001.

[83] Anthony J Viera and Stacey L Sheridan. Global risk of coronary heart disease: assessment and application. *Am Fam Physician*, 82(3):265–274, 2010.

[84] Rod Jackson, Carlene MM Lawes, Derrick A Bennett, Richard J Milne, and Anthony Rodgers. Treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk. *The Lancet*, 365(9457):434–441, 2005.

[85] Neil J Stone, C Noel Bairey Merz, FACC ScM, FAHA Conrad B Blum, FAHA Patrick McBride, FAHA Robert H Eckel, FAHA J Sanford Schwartz, Anne C Goldberg, and FAHA Susan T Shero. 2013 acc/aha guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults. 2013.

[86] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel. Cardiovascular disease risk profiles. *Am Heart J*, 121(1):293–8, January 1991. Available at http://www.ncbi.nlm.nih.gov/pubmed/1985385.

[87] Peter W. F. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, and William B. Kannel. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, 97(18):1837–1847, 1998. Available at http://circ.ahajournals.org/content/97/18/1837.abstract.

[88] R.B. D'Agostino Sr., R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, and W.B. Kannel. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*, 117(6):743–753, 2008. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-39549093148&partnerID=40&md5=5f06ec5a5a99e91d06ba48e365745a4b.

[89] Stephen John Walters. *What is a Cox model?* Available at http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/cox_model.pdf.

[90] G.C.M. Siontis, I. Tzoulaki, K.C. Siontis, and J.P.A. Ioannidis. Comparisons of established risk prediction models for cardiovascular disease: Systematic review. *BMJ (Online)*, 344(7859), 2012. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-84861842176&partnerID=40&md5=b7af5271dde7b6d65c9e97ca9d7dd7b6.

[91] National Institute for Health and UK Care Excellence. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. July 2014 (last accessed: 18-10-15). Available at https://www.nice.org.uk/guidance/cg181/chapter/1-recommendations.

[92] R.M.Conroy et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur. Heart J.*, 24(11):987–1003, 2003. Available at http://eurheartj.oxfordjournals.org/content/24/11/987.abstract.

[93] Ian Graham, Dan Atar, Knut Borch-Johnsen, Gudrun Boysen, Gunilla Burell, Renata Cifkova, Jean Dallongeville, Guy De Backer, Shah Ebrahim, Bjørn Gjelsvik, et al. European guidelines on cardiovascular disease prevention in clinical practice: executive summary. *European heart journal*, 28(19):2375–2414, 2007.

[94] M. Woodward, P. Brindle, and H. Tunsfall-Pedoe. Adding social deprivation and family history to cardiovascular risk assessment: The ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*, 93(2):172–176, 2007.

[95] S Van Dieren, JWJ Beulens, AP Kengne, LM Peelen, GEHM Rutten, M Woodward, YT Van der Schouw, and KGM Moons. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart*, 98(5):360–369, 2012.

[96] G. Assmann, P. Cullen, and H. Schulte. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Mnster (PROCAM) study. *Circulation*, 105(3):310–315, 2002.

[97] M. T. Cooney, A. Dudina, R. D'Agostino, and I. M. Graham. Risk prediction in cardiovascular medicine: Cardiovascular risk-estimation systems in primary prevention: Do They Differ? Do they make a difference? Can we see the future? *Circulation*, 122:300–10–, 2010. Available at http://www.ncbi.nlm.nih.gov/pubmed/20644026.

[98] P.M. Ridker, J.E. Buring, N. Rifai, and N.R. Cook. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds Risk Score. *J Am Med Assoc*, 297(6):611–619, 2007.

[99] Ralph D'Agostino and William B Kannel. Epidemiological background and design: the framingham study. *Proceedings of the American Statistical Association Sesquicentennial Invited Paper Sessions*, 1989.

[100] Thomas R Dawber, Gilcin F Meadors, and Felix E Moore Jr. Epidemiological approaches to heart disease: The framingham study*. *American Journal of Public Health and the Nations Health*, 41(3):279–286, 1951.

[101] J Martin Bland and Douglas G Altman. Survival probabilities (the kaplan-meier method). *Bmj*, 317(7172):1572–1580, 1998.

[102] Ørnulf Borgan. Nelson–aalen estimator. *Encyclopedia of Biostatistics*, 2005.

[103] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[104] Inger Persson. *Essays on the assumption of proportional hazards in Cox regression*. Acta Universitatis Upsaliensis,, 2002.

[105] John P Klein, Hans C Van Houwelingen, Joseph G Ibrahim, and Thomas H Scheike. *Handbook of survival analysis*. CRC Press, 2013.

[106] Kalimuthu Krishnamoorthy. *Handbook of statistical distributions with applications*. CRC Press, 2006.

[107] K. Gu, C. C. Cowie, and M. I. Harris. Mortality in adults with and without diabetes in a national cohort of the U.S. population, 1971-1993. *Diabetes Care*, 21:1138–45, July 1998. Available at http://www.ncbi.nlm.nih.gov/pubmed/9653609.

[108] J. Cederholm, K. Eeg-Olofsson, B. Eliasson, B. Zethelius, P. M. Nilsson, S. Gudbjrnsdottir, and Swedish National Diabetes Register. Risk prediction of cardiovascular disease in type 2 diabetes: a risk equation from the Swedish National Diabetes Register. *Diabetes Care*, 31:2038–43, 2008. Available at http://www.ncbi.nlm.nih.gov/pubmed/18591403.

[109] R. L. Coleman, R. J. Stevens, R. Retnakaran, and R. R. Holman. Framingham, SCORE, and DECODE risk equations do not provide reliable cardiovascular risk estimates in type 2 diabetes. *Diabetes Care*, 30:1292–3, 2007. Available at http://www.ncbi.nlm.nih.gov/pubmed/17290036.

[110] R. K. Simmons, R. L. Coleman, H. C. Price, R. R. Holman, K. T. Khaw, N. J. Wareham, and S. J. Griffin. Performance of the UK Prospective Diabetes Study Risk Engine and the Framingham Risk Equations in Estimating Cardiovascular Disease in the EPIC- Norfolk Cohort. *Diabetes Care*, 32:708–13, 2009. Available at http://www.ncbi.nlm.nih.gov/pubmed/19114615.

[111] Andrew RH Dalton, Alex Bottle, Michael Soljak, Azeem Majeed, and Christopher Millett. Ethnic group differences in cardiovascular risk assessment scores: national cross-sectional study. *Ethnicity & health*, 19(4):367–384, 2014.

[112] G. S. Collins and D. G. Altman. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ*, 339:b2584–, 2009. Available at http://www.ncbi.nlm.nih.gov/pubmed/19584409.

[113] National Institute for Health (NICE) and Clinical Excellence. Lipid Modification - Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. pages –, 2008.

[114] QRISK benefits from NICE decision. published 8th April 2010, (last accessed: 21-10-15). Available at http://www.ehi.co.uk/news/primary-care/5805.

[115] Shanthi Mendis, Lars H. Lindholm, Simon G. Anderson, Ala Alwan, Rajendra Koju, Basden J.C. Onwubere, Azhar Mahmood Kayani, Nihal Abeysinghe, Alfredo Duneas, Sergo Tabagari, Wu Fan, Nizal Sarraf-Zadegan, Porfirio Nordet, Judith Whitworth, and Anthony Heagerty. Total cardiovascular risk approach to improve efficiency of cardiovascular prevention in resource constrain settings. *Journal of Clinical Epidemiology*, 64(12):1451 – 1462, 2011. Available at http://www.sciencedirect.com/science/article/pii/S0895435611000503.

[116] Clara Kayei Chow, Shanthi Naidu, Krishnam Raju, Rama Raju, Rohina Joshi, David Sullivan, David S Celermajer, and Bruce C Neal. Significant lipid, adiposity and metabolic abnormalities amongst 4535 indians from a developing region of rural andhra pradesh. *Atherosclerosis*, 196(2):943–952, 2008.

[117] Colin Mathers, Doris Ma Fat, and JT Boerma. *The global burden of disease: 2004 update*. World Health Organization, 2008 (last accessed: 20-04-15). Available at http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf.

[118] Abhinav Goyal and Salim Yusuf. The burden of cardiovascular disease in the Indian subcontinent. *Indian Journal of Medical Research*, 124(3):235–244, Sep 2006 (last accessed: 20-04-15). Available at http://www.ncbi.nlm.nih.gov/pubmed/17085827.

[119] S Monira Hussain, Brian Oldenburg, Yuanyuan Wang, Sophia Zoungas, and Andrew M Tonkin. *Assessment of cardiovascular disease risk in south Asian populations*, volume 2013 (last accessed: 20-04-15). Hindawi Publishing Corporation, 2013. Available at http://www.hindawi.com/journals/ijvm/2013/786801/.

[120] Ministry of Health and Family Welfare, Government of India. *National Programme for Prevention and Control of Diabetes, Cardiovascular Disease, and Stroke: A manual for Medical Officer; Government of India-WHO Collaborative Programme 2008-2009*. 2009 (last accessed: 10-12-14). Available at http://gujhealth.gov.in/images/MANUAL_for_medical_officer.pdf.

[121] Gordon P Hughes. On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*, 14(1):55–63, 1968.

[122] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2005.

[123] Sergio Garcia, Julián Luengo, José Antonio Sáez, Victor Lopez, and Francisco Herrera. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *Knowledge and Data Engineering, IEEE Transactions on*, 25(4):734–750, 2013.

[124] Wentian Li. Mutual information functions versus correlation functions. *Journal of statistical physics*, 60(5-6):823–837, 1990.

[125] Athanasios Tsanas. *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning*. PhD thesis, University of Oxford, 2012.

[126] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[127] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM, 2004.

[128] Chih-Jen Lin, Ruby C Weng, and S Sathiya Keerthi. Trust region newton method for logistic regression. *The Journal of Machine Learning Research*, 9:627–650, 2008.

[129] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity: the lasso and generalizations. 2015.

[130] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[131] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[132] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[133] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[134] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[135] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[136] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251. IEEE, 2013.

[137] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. wadsworth. *Belmont, CA*, 1984.

[138] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[139] Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008.

[140] MATLAB implementation of the randomforest R package. Available at https://code.google.com/p/randomforest-matlab/(lastaccessed:01-10-14).

[141] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[142] Leo Breiman. Out-of-bag estimation, 1996. Available at https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf(lastaccessed:01-10-14).

[143] National Institute of Health, Fogarty International Centre. News article -reducing the guesswork and cost of rating heart disease, 2011 (last accessed: 03-12-14). Available at http://www.fic.nih.gov/News/GlobalHealthMatters/Sept-Oct-2011/Pages/chronic-disease-cardiovascular.aspx.

[144] World Health Organization. Health Financing. Available at http://gamapserver.who.int/gho/interactive_charts/health_financing/atlas.html?indicator=i3(lastaccessed: 03-12-14).

[145] World Bank. Total Population of India- World Bank Development Indicators, 2013. Available at http://data.worldbank.org/indicator/SP.POP.TOTL(lastaccessed: 10-12-14).

[146] International Institute for Population Sciences. National Family Health Survey-3, 2007 (last accessed: 10-12-14). Available at http://www.rchiips.org/nfhs/factsheet.shtml.

[147] Shraddha Chauhan and Bani Tamber Aeri. Prevalence of cardiovascular disease in India and it is economic impact-a review. *International Journal of Scientific and Research Publications*, 3:212, 2013.

[148] Jinchuan Xing, W Scott Watkins, Ya Hu, Chad Huff, Aniko Sabo, Donna Muzny, Michael Bamshad, Richard Gibbs, Lynn Jorde, and Fuli Yu. Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biology*, 11(11):R113, 2010. Available at http://genomebiology.com/2010/11/11/R113.

[149] Mangesh S Pednekar, Rajeev Gupta, and Prakash C Gupta. Association of blood pressure and cardiovascular mortality in india: Mumbai cohort study. *American journal of hypertension*, 22(10):1076–1084, 2009.

[150] R Gupta, RM Pandey, A Misra, A Agrawal, P Misra, S Dey, S Rao, VU Menon, N Kamalamma, KP Vasantha Devi, et al. High prevalence and low awareness, treatment and control of hypertension in asian indian women. *Journal of human hypertension*, 26(10):585–593, 2012.

[151] Blaž Zupan, Janez DemšAr, Michael W Kattan, J Robert Beck, and Ivan Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1):59–75, 2000.

[152] Michael W Kattan. Comparison of cox regression with other methods for determining prediction models and nomograms. *The Journal of urology*, 170(6):S6–S10, 2003.

[153] Brian D Ripley and Ruth M Ripley. Neural networks as statistical methods in survival analysis. *Clinical applications of artificial neural networks*, pages 237–255, 2001.

[154] Leonardo Vanneschi, Antonella Farinaccio, Giancarlo Mauri, Mauro Antoniotti, Paolo Provero, and Mario Giacobini. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData mining*, 4(1):1, 2011.

[155] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, and Honglak Lee. An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 183–192. ACM, 2010.

[156] Gustavo EAPA Batista and Maria Carolina Monard. A study of k-nearest neighbour as an imputation method. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.3558&rep=rep1&type=pdf.

[157] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. Feature extraction: foundations and applications. 207, 2008.

[158] Hervé Stoppiglia, Gérard Dreyfus, Rémi Dubois, and Yacine Oussar. Ranking a random feature for variable and feature selection. *The Journal of Machine Learning Research*, 3:1399–1414, 2003.

[159] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256, 1992.

[160] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[161] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[162] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[163] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4), 2011.

[164] Tao Li, Shenghuo Zhu, and Mitsunori Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and information systems*, 10(4):453–472, 2006.

[165] VF Rodriguez-Galiano, B Ghimire, J Rogan, M Chica-Olmo, and JP Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104, 2012.

[166] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[167] Stanley Lemeshow and David W Hosmer. A review of goodness of fit statistics for use in the development of logistic regression models. *American journal of epidemiology*, 115(1):92–106, 1982.

[168] Andrew A Kramer and Jack E Zimmerman. Assessing the calibration of mortality benchmarks in critical care: The hosmer-lemeshow test revisited*. *Critical care medicine*, 35(9):2052–2056, 2007.

[169] James P Marcin and Patrick S Romano. Size matters to a model's fit*. *Critical care medicine*, 35(9):2212–2213, 2007.

[170] WHO Expert Consultation. Waist circumference and waist-hip ratio. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.418.302&rep=rep1&type=pdf.

[171] V Mohan and R Deepa. Obesity & abdominal obesity in asian indians. *Indian Journal of Medical Research*, 123:593–596, 2006.

[172] Nick Freemantle, Jeremy Holmes, A Hockey, and Sudhesh Kumar. How strong is the association between abdominal obesity and the incidence of type 2 diabetes? *International journal of clinical practice*, 62(9):1391–1396, 2008.

[173] FE Von Eyben, E Mouritsen, J Holm, P Montvilas, G Dimcevski, G Suciu, I Helleberg, L Kristensen, and R Von Eyben. Intra-abdominal obesity and metabolic risk factors: a study of young adults*. *International journal of obesity*, 27(8):941–949, 2003.

[174] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. pages 161–168, 2006.

[175] David J Hand et al. Classifier technology and the illusion of progress. *Statistical science*, 21(1):1–14, 2006.

[176] Lawrence de Koning, Anwar T Merchant, Janice Pogue, and Sonia S Anand. Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: meta-regression analysis of prospective studies. *European heart journal*, 28(7):850–856, 2007.

[177] Luis Alcocer, Mario Bendersky, Julio Acosta, and Miguel Urina-Triana. Use of calcium channel blockers in cardiovascular risk reduction. *American journal of cardiovascular drugs*, 10(3):143–154, 2010.

[178] Cristina Bianchi, Roberto Miccoli, Giuseppe Daniele, Giuseppe Penno, and Stefano Del Prato. Is there evidence that oral hypoglycemic agents reduce cardiovascular morbidity/mortality? yes. *Diabetes care*, 32(suppl 2):S342–S348, 2009.

[179] Amber AWA van der Heijden, Monica M Ortegon, Louis W Niessen, Giel Nijpels, and Jacqueline M Dekker. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: Accuracy of the framingham, score, and ukpds risk functions the hoorn study. *Diabetes care*, 32(11):2094–2098, 2009.

[180] Lourdes Cañón Barroso, Eloísa Cruces Muro, Natalio Díaz Herrera, Gerardo Fernández Ochoa, Juan Ignacio Calvo Hueros, and Francisco Buitrago. Performance of the framingham and score cardiovascular risk prediction functions in a non-diabetic population of a spanish health care centre: a validation study. *Scandinavian journal of primary health care*, 28(4):242–248, 2010.

[181] Nicholas J Wald, Mark Simmonds, and Joan K Morris. Screening for future cardiovascular disease using age alone compared with multiple risk factors and age. *PLoS ONE*, 6(5), 2011.

[182] Richard Smith. Screening for cardiovascular disease using age alone: reflections on a paper peer-reviewed as both 'radical'and 'unsurprising'. *Journal of medical screening*, 18(3):113–114, 2011.

[183] The Indian Polycap Study. Effects of a polypill (polycap) on risk factors in middle-aged individuals without cardiovascular disease (tips): a phase ii, double-blind, randomised trial. *The Lancet*, 373(9672):1341–1351, 2009.

[184] Eva Lonn, Jackie Bosch, Koon K Teo, Prem Pais, Denis Xavier, and Salim Yusuf. The polypill in the prevention of cardiovascular diseases key concepts, current status, challenges, and future directions. *Circulation*, 122(20):2078–2088, 2010.

[185] Karel GM Moons, Douglas G Altman, Yvonne Vergouwe, Patrick Royston, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj*, 338:b606, 2009.

[186] Andrew Beswick and Peter Brindle. Risk scoring in the assessment of cardiovascular risk. *Current opinion in lipidology*, 17(4):375–386, 2006.

[187] Mohsen Aarabi and Peter R Jackson. Predicting coronary risk in uk south asians: an adjustment method for framingham-based tools. *European Journal of Cardiovascular Prevention & Rehabilitation*, 12(1):46–51, 2005.

[188] Paulino González-Diego, Conchi Moreno-Iribas, María Jesús Guembe, José Javier Viñes, and Joan Vila. Adaptation of the framingham-wilson coronary risk equation for the population of navarra (ricorna). *Revista Española de Cardiología (English Edition)*, 62(8):875–885, 2009.

[189] C.K. Chow, R. Joshi, D.S. Celermajer, A. Patel, and B.C. Neal. Recalibration of a framingham risk equation for a rural population in india. *Journal of Epidemiology and Community Health*, 63(5):379–385, 2009. Available at http://www.scopus.com/inward/record.url?eid=2-s2.0-66149126633&partnerID=40&md5=117c686b51ef1e80c3d71a75169de584.

[190] Karen A Matthews, Mary Fran Sowers, Carol A Derby, Evan Stein, Heidi Miracle-McMahill, Sybil L Crawford, and Richard C Pasternak. Ethnic differences in cardiovascular risk factor burden among middle-aged women: Study of women's health across the nation (swan). *American heart journal*, 149(6):1066–1073, 2005.

[191] Salim Yusuf, Sumathy Rangarajan, Koon Teo, Shofiqul Islam, Wei Li, Lisheng Liu, Jian Bo, Qinglin Lou, Fanghong Lu, Tianlu Liu, et al. Cardiovascular risk and events in 17 low-, middle-, and high-income countries. *New England Journal of Medicine*, 371(9):818–827, 2014.

[192] Timo E Strandberg and Kaisu Pitkala. What is the most important component of blood pressure: systolic, diastolic or pulse pressure? *Current opinion in nephrology and hypertension*, 12(3):293–297, 2003.

[193] Stanley S Franklin, Victor A Lopez, Nathan D Wong, Gary F Mitchell, Martin G Larson, Ramachandran S Vasan, and Daniel Levy. Single versus combined blood pressure components and risk for cardiovascular disease the framingham heart study. *Circulation*, 119(2):243–250, 2009.

[194] Giuseppe Schillaci, Matteo Pirro, and Elmo Mannarino. Assessing cardiovascular risk should we discard diastolic blood pressure? *Circulation*, 119(2):210–212, 2009.

[195] Tanu Midha, Arati Lalchandani, Bhola Nath, Ranjeeta Kumari, and Umeshwar Pandey. Prevalence of isolated diastolic hypertension and associated risk factors among adults in kanpur, india. *Indian heart journal*, 64(4):374–379, 2012.

[196] S. Theodoridis and K Koutroumbas. *Pattern Recognition (3rd edition)*. Academic Press, 2006.

[197] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*, volume 20. Siam, 2007.

[198] David Wishart. K-means clustering with outlier detection, mixed variables and missing values. In *Exploratory Data Analysis in Empirical Research*, pages 216–226. Springer, 2003.

[199] M. Ichino and H. Yaguchi. Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(4):698–708, April 1994. Available at http://dx.doi.org/10.1109/21.286391.

[200] Fionn Murtagh and Pierre Legendre. Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. *arXiv preprint arXiv:1111.6285*, 2011.

[201] A.D Gordon. *Classification*. Chapman Hall/CRC, Boca Raton, FL, 2nd edition, 1999.

[202] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, apr 1979. Available at http://dx.doi.org/10.1109/tpami.1979.4766909.

[203] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[204] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[205] Anil K Jain. Data clustering: 50 years beyond k-means.

[206] MATLAB and Statistics Toolbox Release. *version 8.3.0 (R2014a)*. The MathWorks Inc., Natick, Massachusetts, 2014.

[207] Karla Hemming, Richard Lilford, and Alan J Girling. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Statistics in medicine*, 34(2):181–196, 2015.

[208] Marlies Noordzij, Giovanni Tripepi, Friedo W Dekker, Carmine Zoccali, Michael W Tanck, and Kitty J Jager. Sample size calculations: basic principles and common pitfalls. *Nephrology dialysis transplantation*, page gfp732, 2010.

[209] Michael A Hussey and James P Hughes. Design and analysis of stepped wedge cluster randomized trials. *Contemporary clinical trials*, 28(2):182–191, 2007.

[210] Christian Allena'b, Darius Jazayeri, Justin Miranda, Paul G Biondich, Burke W Mamlinc, Ben A Wolfe, Chris Seebregtsd, Neal Lesha, William M Tierney, and Hamish SF Frasera'e. Experience in implementing the openmrs medical record system to support. *Medinfo 2007*, page 382, 2007.

[211] Christopher J Seebregts, Burke W Mamlin, Paul G Biondich, Hamish SF Fraser, Benjamin A Wolfe, Darius Jazayeri, Christian Allen, Justin Miranda, Elaine Baker, Nicholas Musinguzi, et al. The openmrs implementers network. *International journal of medical informatics*, 78(11):711–720, 2009.

[212] OpenMRS Developers Guide, FLOSS manuals. 2015 (last accessed: 16-09-15). Available at https://flossmanuals.net/openmrs-developers-guide/.

[213] Andrew J Farmer, Oliver J Gibson, Christina Dudley, Kathryn Bryden, Paul M Hayton, Lionel Tarassenko, and Andrew Neil. A randomized controlled trial of the effect of real-time telemedicine support on glycemic control in young adults with type 1 diabetes (isrctn 46889446). *Diabetes care*, 28(11):2697–2702, 2005.

[214] Microsoft TechNet. What is tls/ssl? 2015 (last accessed: 25-10-15).

[215] Peter Byass, Daniel Chandramohan, Samuel J Clark, Lucia D'Ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, et al. Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global health action*, 5, 2012.

[216] Christine Zhenwei Qiang, Masatake Yamamichi, Vicky Hausman, Daniel Altman, and ICT Sector Unit. Mobile applications for the health sector. *Washington: World Bank*, 2011.

[217] Nisha I Parikh, Michael J Pencina, Thomas J Wang, Emelia J Benjamin, Katherine J Lanier, Daniel Levy, Ralph B D'Agostino, William B Kannel, and Ramachandran S Vasan. A risk score for predicting near-term incidence of hypertension: the framingham heart study. *Annals of internal medicine*, 148(2):102–110, 2008.

[218] Patrick Mair and Jan de Leeuw. A general framework for multivariate analysis with optimal scaling: The r package aspect. *Department of Statistics, UCLA*, 2009.

Appendix A

Worked examples of point-of-care algorithm in Chapter 5

We present two worked examples of using the Point-of-care (POC) simplified test for determining if a total cholesterol (TC) test will benefit CVD risk prediction using the WHO/ISH charts. Patient data is taken from our testing set of the APHRI dataset.

Example 1

Consider a 70 year old male patient whose data is being collected at point of care. He is a non-smoker and weighs 49kg and is 162.5 cm tall. The patient has two consecutive blood pressure measurements taken which are recorded to be 137/87 mmHg and 138/84 mmHg. He also records a random blood glucose level of 79 mg/dl.

1. Prior to determining 10-year CVD risk, we can use equation (1) as described in the paper,
$$logit(c) = -5.6554 + 0.0416 * Age + 0.0132 * SBP \tag{A.1}$$
where c is the probability for a patient to require a cholesterol test.

2. The patient's mean SBP is calculated to be 137.5 mmHg.
$$\begin{aligned} logit(c_1) &= -5.6554 + 0.0416 * (70) + 0.0132 * (137.5) \\ c_1 &= \frac{1}{(1 + e^{-0.9284})} \\ c_1 &= 0.2832 \end{aligned}$$

3. The threshold corresponding to the maximum F3 score, which has a high sensitivity, is 0.1215. Alternatively, the threshold corresponding to the F1 measure may also be considered (0.1762) if the user desires a balance between sensitivity and specificity. Both thresholds are below c_1 indicating that a TC test will be beneficial for risk estimation which implies that it is preferable to use the WHO/ISH HI charts.

The patient's complete medical data reveals a TC measurement to be 221 mg/dl. Hence the 10-year risk computed using the WHO/ISH LI charts would have been 20 to <30% risk while risk computed using the HI charts would be 30 to <40% risk. As a side note, if the patient had been assessed for risk only using the WHO/ISH LI charts, the NPCDCS guidelines would not have classified the patient as requiring treatment (10-year CVD risk would have been between 20 to <30% and SBP<140mmHg). On the other hand, the patient's risk according to the HI charts clearly mandates that he ought to be high risk and requiring treatment.

Example 2

Consider a 48-year old male patient, non-smoker with height 158cm and weight 61kg. He has a mean BP of 154/89 mmHg from two consecutive readings and a glucose level (random) of 110 mg/dl.

1. Using equation (1), we find

$$\begin{aligned} logit(c_2) &= -5.6554 + 0.0416 * (48) + 0.0132 * (154) \\ c_2 &= \frac{1}{(1+e^{-1.6258})} \\ c_2 &= 0.1644 \end{aligned}$$

2. If we choose the threshold corresponding to the maximum F3 score (0.1215), a TC test and risk estimation by the WHO/ISH HI charts are recommend. However, the threshold for maximum F1 score (0.1762) which offers higher specificity would say the TC test won't be necessary and WHO/ISH LI charts would suffice.

The patient's complete medical data tells us his recorded TC was 208 mg/dl. Both the WHO/ISH LI charts and HI charts would estimate the 10-year risk to be less than 10% risk.

The variation of age, SBP, and c is shown through Fig. A.1.

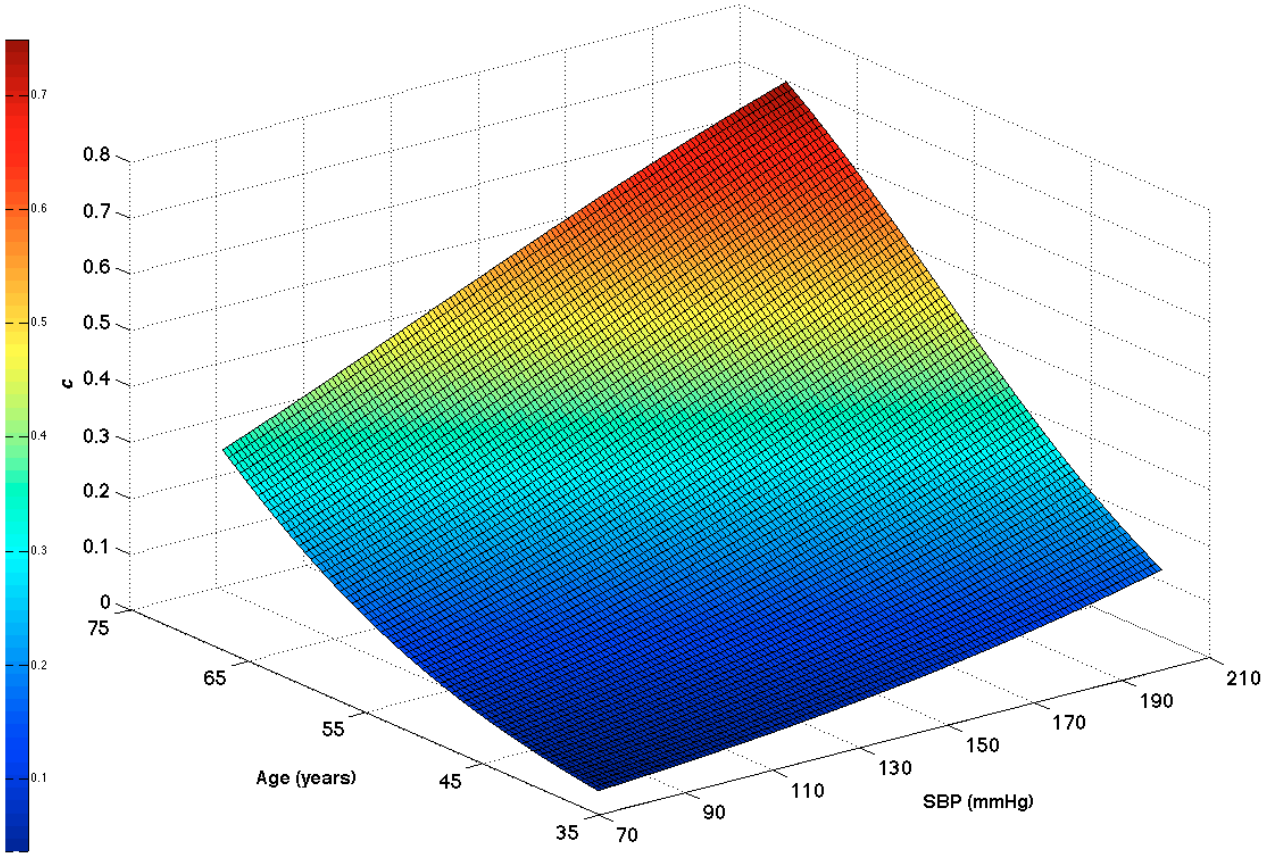


Figure A.1 The variation of c with age and systolic blood pressure.

Appendix B

Historical event list referenced in Chapter 2

Figure B.1 List of historical events incorporated in the mobile CDSS to obtain accurate estimate of age. The diagram illustrates the example in case of a male participant.

Male event list

- 1947: India Independence day
- 1948: Assassination of Mahatma Gandhi

- 1948: Liberation of Hyderabad from Nizam's Rule
- 1950: India's Republic day
- 1956: Formation of Andhra Pradesh state
- 1962: India-China war
- 1969: Man landed on Moon first time
- 1970: War with Pakistan
- 1971: Jai Andhra Movement
- 1975-1976: Emergency
- 1977: Janatha party came into power
- 1977: Big cyclone and Diviseema floods
- 1979: Fall of Skylab
- 1983: Formation of TDP government by N T Rama Rao
- 1984: Assassination of Prime Minister Indira Gandhi
- 1985: Rajiv Gandhi became Prime Minster of India
- 1986: Big floods in coastal Andhra
- 1988: Assassination of famous politician V V M Ranga

Female event list

Option A For married women having no children, the following questions may be asked

- Age at puberty (if unknown, record as 13 years)
- Duration between puberty and first marriage
- Duration of the first marriage
- To obtain present age, add all of the above responses

Option B For married women with at least one child, the following questions may be asked

- Age at puberty (if unknown, record as 13 years)
- Duration between puberty and first marriage
- Duration between first marriage and first child

- Present age of the first child
- To obtain present age, add all of the above responses

Appendix C

Barriers to adoption of mHealth system in Chapter 2

The following content is extracted from a book chapter [51] written by the author of this thesis that narrates lessons learnt from the evaluation of the mHealth CDSS, *SMARThealth* in rural India.

General user acceptance of the CDSS Qualitative findings indicated that all ASHAs found the CDSS useful and felt that not only did it enhance their capacity as health workers, but also helped in raising patient awareness of CVD and healthy behaviour.

"I like all the features of the tool. I feel it will be very useful to the community by creating awareness about their health and to take care of it. This tool is giving an opportunity to them to learn about their health." (ASHA 6, age 29)

"They (patients screened) all felt happy to know their CVD risk in advance and to learn about the impending risk of those diseases. They also informed that by knowing about these things, what care needs to be taken is made clear to them. They are all happy about the utility of this tablet and the screening procedure" (ASHA 2, age 29)

The CDSS also benefitted the ASHAs over conventional pa-per-based data collection methods. The use of the Bluetooth blood pressure device, and glucometers accelerated the process of collecting risk factor measurements.

"If paper based, it would take a lot of time, it may take nearly half an hour to record everything But in TAB, it takes a some time at the registration page but takes no time to measure BP, sugar etc., if we do it quickly, it would end fast." (ASHA 9, age 31)

Adoption of technology During the UCD phase, different ways of communicating the meaning of absolute CVD risk developed over time (such as through graphs) was tested on participants from different villages. The final design of the CDS based CVD risk tool used a risk projection 'meter' using a needle and dial, similar to a speedometer, with the risk ranging from 0 to 100%, with the addition of colour coding which mapped to the WHO/ISH charts. The ASHAs found this useful to explain the relation between different risk factors such as BP and absolute risk.

"We show them (patients) the risk meter. When age is moved, this bar moves along. When the age or BP is high this increases so, we are showing this....This is easier to understand, as this will move from here to here, whereas that doesn't move, so this is easier to understand and also to show the difference comparing with the age and BP" (ASHA 10, age 30)

"I used to show them (patients) this (risk meter) and let them know about the BP.... Simply explaining will not work. If we show this page they can understand it well. If we show them what is risk, what is normal reading, they understand it well" (ASHA 9, age 31)

The ASHAs experienced difficultly in operating the Bluetooth BP device initially, and some preferred manual entry or sought assistance. Towards the later stages of the pilot study (which was spread over 4 weeks), they reported less difficulty and ex-pressed a preference for Bluetooth transmission to collect the BP data.

"In the beginning I could not have the reading afterwards, it is ok... First it was difficult and now it is easy to transfer. If I press the button, it will take the reading" (ASHA 9, age 31)

"Yes sir, once I felt difficulty in the beginning and then I called sir over phone and asked for the guidance... Initially I thought typing BP values are easy..I felt somewhat difficulty (to send BP data wirelessly) in the beginning....afterwards, I could do it... " (ASHA 11, age 35)

Barriers to increased adoption Use of pictograms was recommended by the ASHAs to make a greater impact when communicating lifestyle modifications such as salt reduction and physical activity to local members of the community.

"That will still be better and if we show the spoon in a salt bowel or pictures of exercise and ... if we show the pictures it will be better... and they will be seen and hence it is easy; to explain also... and they understand by seeing... " (ASHA 4, age 32)

Although members of the community valued the role of the CVD risk projection dial in understanding the relation between different risk factors and absolute CVD risk, they did not

perceive the meaning of absolute risk in itself. This indicated the need for better mechanisms of conveying absolute risks, such as through coloured bars and the need to minimise patient exposure to their percentage of CVD risk.

"They (patients) don't understand the percentages....I would tell them that as you have high risk you have to go to the doctor, or if the risk is too low then would say you won't need a doctor as your risk is very low. Some of them are asking and some of them are not asking such questions on what is this percentage. I suggest without using the words percentage we can convey them this in degrees like more, less and normal or too high" (ASHA 10, age 30)

The low-cost Android tablets used by the ASHAs encountered difficulties including the uptake of power when connected to a mains power supply. This was compounded by the limited avail-ability of electricity throughout the day in rural India.

"Charging is a main issue for me. My tablet was not getting charged though I have been putting it on charge for a long time"(ASHA 6, age 29)

During the training phase of the study, the introduction of auto-mated text messages from the OpenMRS electronic health record system found poor acceptability amongst community participants. This was because a majority of eligible participants did not read text messages (which were in English and added to the language barrier). They were furthermore used to receiving frequent spam texts. The SMS feature was therefore found infeasible and was dropped for the pilot study.

Appendix D

List of features selected in Chapter 5

The following set of 40 features that have been reported in CVD literature were chosen from the APRHI dataset for input into a feature selection algorithm.

- Gender
- Age
- Educational level
- Current smoker
- Cigarettes per day
- Number of smokers in household
- Past history of heart attack
- Past history of angina
- Past history of stroke
- Past history of peripheral vascular disease
- Past history of cardiovascular disease
- Past history of diabetes
- Past history of hypertension
- Past history of elevated total cholesterol
- Relatives diagnosed with diabetes
- Relatives diagnosed with heart attack before 60 years

- Relatives diagnosed with stroke before 60 years
- Height
- Weight
- Body-mass index
- First systolic blood pressure measurement
- First diastolic blood pressure measurement
- First heart rate measurement
- Second systolic blood pressure measurement
- Second diastolic blood pressure measurement
- Second heart rate measurement
- Average systolic blood pressure measurement
- Average diastolic blood pressure measurement
- Average heart rate measurement
- Glucose
- Fasting before glucose measurement
- Total cholesterol
- HDL cholesterol
- LDL cholesterol
- Triglycerides
- Family history of CVD
- Medication - ACE inhibitors
- Medication - Beta blockers
- Medication - Calcium Antagonists
- Medication - Diuretics
- Medication - treatment for BP
- Medication - Statin